

A SOCIO-TECHNICAL APPROACH TO PROTECTING PEOPLE'S
PRIVACY IN THE CONTEXT OF SHARING IMAGES ON SOCIAL
MEDIA

Rakibul Hasan

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics, Computing, and Engineering,
Indiana University
December, 2020

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

Doctoral Committee

Apu Kapadia, PhD, Chair

David J. Crandall, PhD

Bennett I. Bertenthal, PhD

Sameer Patil, PhD

September 24, 2020

Copyright © 2020

Rakibul Hasan

Dedicated to my wife,

Jhumu

Her endless support and countless sacrifices made this journey possible.

ACKNOWLEDGMENTS

This dissertation would not have been possible without enormous help from my advisor, research committee members, colleagues, friends, and family members. I would like to take this opportunity to express my heartfelt gratitude to all these people.

First, I would like to express my appreciation to my advisor Dr. Apu Kapadia, for believing in me, for his constant encouragement, support, and guidance. The Ph.D. is a long and stressful journey with lots of ups and downs, but Apu made it easier to navigate with his persistent care for my well-being. Apu's prudent guidance helped improve my personal life in many ways, for which I will be forever grateful.

I would also like to thank my committee members for their guidance and useful comments that helped to improve my dissertation. I am thankful to Dr. David Crandall and Dr. Bennett Bertenthal for their collaboration on my dissertation research. I would also like to thank Dr. Sameer Patil for his insightful comments on drafts of my dissertation. My research was partly funded by the US National Science Foundation (NSF) through grants CNS-1408730, IIS-1253549, and IIS-1527421; I am thankful to NSF.

From the beginning of my Ph.D. study, I have been collaborating with Dr. Roberto Hoyle, who was also a member of the IU Privacy lab and is now at Oberlin College, and Dr. Kelly Caine at Clemson University. For the past several years, I have been collaborating with Dr. Kurt Hugenberg at the Dept. of Psychology and Brain Sciences at Indiana University, and Dr. Mary Jean Amon, currently at the University of Central Florida. Their insightful guidance enabled me to successfully complete those projects and publish research findings. I have been fortunate to collaborate with Dr. Yong-Yeol Ahn; although this research was outside the scope of my dissertation, I gained new insights into both research methodology and statistical analyses that will help me in my future research. Under this project, I also collaborated with Dr. Cristobal Cheyre at Cornell University. I

am thankful to all of them for sharing their knowledge and expertise to help me succeed. I also want to express my gratitude to Eman Hassan and Yifang Li, with whom I collaborated and co-authored multiple papers. I am also grateful to Dr. Shawn Fagan for always offering me valuable feedback on manuscripts and presentations.

I am very grateful to all past and present members of the IU Privacy Lab, with whom I shared so many happy moments. I have collaborated with Dr. Tousif Ahmed and Taslima Akter, and learned a lot from them. In particular, Tousif helped me in numerous situations, both academic and non-academic.

I am greatly indebted to my wife, Jhumu, for her never-ending support and encouragement throughout this long, challenging journey. I will be forever indebted to my parents for their unconditional sacrifices throughout my life. I come from a family with low socio-economic status; my parents did their best and made countless sacrifices so that I could have a better life than they did. Their dreams inspired me to take on and continue this journey, and will provide me the courage to face any future challenges.

Rakibul Hasan

A SOCIO-TECHNICAL APPROACH TO PROTECTING PEOPLE'S PRIVACY IN THE
CONTEXT OF SHARING IMAGES ON SOCIAL MEDIA

Billions of photos are being shared on social media platforms every day. A large portion of these photos are taken in public places, and may contain people who were inadvertently captured (i.e., bystanders) and are not important for the subject matter of the photos. When these photos are shared online, they reveal the bystanders' identity, location, and other privacy-sensitive information to a potentially unbounded number of internet users. Social media users not only share photos they own but also re-share photos from their peers and those they find on the internet; for example, the sharing of image macros or memes on social media has risen in popularity. Internet users create memes using photos of other people who are often unknown to them. Such photos usually portray people in embarrassing situations, which are highlighted and amplified with additional text or captions. These photos can go 'viral' and cause severe personal, social, and professional consequences to the photo subjects. While the research community has made significant efforts to reduce photo-sharers' privacy risks on social media, protecting the privacy of people who do not actively take part in photo-taking or sharing activities, e.g., bystanders and meme subjects, has not received adequate attention. This dissertation proposes machine learning and computer vision-based techniques to reduce bystanders' privacy risks. More specifically, we offer an automated and scalable system to detect bystanders in images so that their privacy can be protected by, e.g., removing or obfuscating them using image transforms. In an online study, we evaluated the effectiveness and usability of commonly used image transforms. We constructed and empirically validated models of interactions among image filters and utility variables. Based on these models, we proposed a principled approach to design novel obfuscations to balance the privacy-utility trade-offs. To protect the privacy of meme subjects, we explored the potential of behavioral interventions to discourage

meme sharing. Through controlled experiments, we identified demographic factors and personality traits that affect behaviors regarding photo sharing that may threaten other people's privacy. We also discovered links between people's personality traits and their reactions to privacy nudges that were designed to discourage them from sharing memes. These results can be used to develop direct and personalized interventions to stimulate privacy-respecting and prosocial behaviors among social media users.

Apu Kapadia, PhD, Chair

David J. Crandall, PhD

Bennett I. Bertenthal, PhD

Sameer Patil, PhD

CONTENTS

1	Introduction	1
2	Problem and Thesis Statement	7
2.1	Thesis	8
2.1.1	Dissertation Outline	10
2.1.2	Designing Machine Learning-Based Models to Automatically Classify ‘By-stander’ and ‘Subject’ in Images	11
2.2	Evaluating the Privacy-Protection Capability and Usability of Image Filters	12
2.3	Designing Novel Image Obfuscations	13
2.4	Individual Differences and Photo-sharing Behaviors	14
3	Related Work	16
3.1	Sharing Photos on Social Media	16
3.2	Privacy Risks in the Context of Sharing Photos on Social Media	16
3.3	Protecting Bystanders’ Privacy	18
3.3.1	Privacy Protection in the Moment of Photo Capture	18
3.3.2	Protecting Bystanders’ Privacy in Images in the Cloud	20
3.4	Approaches to Reducing Privacy Risks	21
3.4.1	Limiting Dissemination	21
3.4.2	Reducing Privacy Risks by Obfuscating Sensitive Scene Elements	22
3.5	Employing Behavioral Interventions to Reduce Privacy Risks	24
3.5.1	Effects of Personality Traits on Photo-sharing Behaviors	25
3.5.2	Demographic Differences in Photo-Sharing Behaviors	25
3.5.3	Personalized Interventions	26

4 Machine Learning Based Models for Automatic Classification of ‘Bystander’ and ‘Subject’ in Photos	28
4.1 Introduction	28
4.2 Study Method	31
4.2.1 Survey Design	33
4.2.2 Survey Participants and Dataset Labels	37
4.3 Method of Analysis	39
4.3.1 Quantifying association between human reasoning and features	39
4.3.2 Measuring predictive-power of individual feature and selecting subset of uncorrelated features	40
4.3.3 Developing classifiers using selected feature sets	41
4.3.4 Comparing ML models with humans	44
4.3.5 Test Dataset	45
4.4 Findings	46
4.4.1 How Humans Classify ‘Subjects’ and ‘Bystanders’?	46
4.4.2 Association Between Human-reasoning and the Features	47
4.4.3 Machine Learning Models to Predict ‘Subject’ and ‘Bystander’	55
4.4.4 Comparing ML Models with Humans	56
4.4.5 Accuracy on the COCO Dataset	56
4.5 Limitations and Discussion	57
4.6 Conclusion	59
5 Evaluating Image Filters in Terms of Privacy-Protection Capability and Usability	61
5.1 Introduction	61
5.2 Experiment	62

5.2.1	Measurements	63
5.2.2	Scene selection	65
5.2.3	Obfuscation Methods	66
5.2.4	Collecting Images	67
5.2.5	Organization of the Survey	68
5.2.6	Ethical Considerations	69
5.2.7	Recruitment, Compensation, and Validation	69
5.2.8	Pilot Study	69
5.3	Findings	70
5.3.1	Demographic Information	70
5.3.2	Recognition Accuracy	70
5.3.3	Recognition Confidence	73
5.3.4	Photo Utility	73
5.3.5	Privacy-Utility Trade-off	76
5.4	Discussion	81
5.4.1	Privacy vs. Utility	81
5.4.2	Effectiveness of Filters Throughout Categories	81
5.4.3	Edge Detection Side Effects	82
5.4.4	Implications and Practical Applications	82
5.4.5	Human vs. Computer Viewers	83
5.4.6	Limitations	84
5.5	Conclusions	84
6	Designing Novel Image Obfuscations	88
6.1	Introduction	88
6.2	Method	90

6.2.1	Path Model Analysis	91
6.2.2	Path Model Results	92
6.2.3	Experimental Design	95
6.2.4	Participants	95
6.2.5	Selecting Attributes	96
6.2.6	Image dataset	97
6.2.7	Privacy-enhancing Transformations and Artistic Transformations	97
6.2.8	Measurements	99
6.2.9	Procedure	100
6.2.10	Data Analysis Procedure	100
6.3	Findings	102
6.3.1	Demographic Characteristics of the Participants	102
6.3.2	Effects of Transformations on Information Content	102
6.3.3	Effects of Transformations on Visual Aesthetics	103
6.3.4	Effects of Transformations on Viewers' Satisfaction	105
6.4	Discussion	108
6.4.1	Limitations	109
6.5	Conclusions	110
7	Individual Differences and Photo-sharing Behaviors	112
7.1	Introduction	113
7.2	Background	114
7.2.1	Internet Memes and their Functions	115
7.2.2	Individual Humor Style	115
7.2.3	Relevance of 'Humor style' to Photo-sharing Behaviors	116
7.2.4	Justifications of the Interventions	117

7.3	Method	119
7.3.1	Collecting Memes	119
7.3.2	Study I: Collecting Valence Ratings of the Memes	120
7.3.3	Study II: Collecting Data on Photo-sharing Behaviors	122
7.3.4	Methods of Data Analysis	127
7.4	Findings	129
7.4.1	Relation Between Humor Type and Photo-sharing Behaviors	129
7.4.2	Reactions to the Interventions	132
7.4.3	Effect of Gender	133
7.5	Discussion	135
7.5.1	<i>Humor Endorsers</i> are More Likely to Share Memes with <i>Very Negative</i> Valence.	136
7.5.2	Reactions to the <i>Perspective Taking</i> Intervention.	136
7.5.3	Reactions to the <i>Privacy Perspective</i> Intervention.	137
7.5.4	Effects of Gender.	139
7.5.5	Effect of Time Delay.	140
7.5.6	Limitations	140
7.6	Conclusions	141
8	Discussions, Limitations, and Future Work	145
8.1	Limitations and Future Work	147
8.1.1	Social Relationships were not used when Classifying Bystanders and Subjects in Photos	147
8.1.2	Image Obfuscations' Acceptability was Studied from Photo-Viewers' Perspec- tive	147
8.1.3	Unintended Consequences of Image Obfuscations	147
8.1.4	Improving the Classification Accuracy of the Bystander Detection Model . .	148

8.1.5	Designing Better Image Obfuscations	149
8.1.6	Visual Interventions to Discourage the Sharing of Privacy-Sensitive Photos .	150
8.1.7	Evaluating ML Models, Obfuscation Methods, and Behavioral Interventions in the Wild.	151
9	Conclusions	153
	References	155
	Appendices	181
	Curriculum Vitae	

LIST OF FIGURES

4.1	Example stimuli used in our survey.	34
4.2	Detecting and refining body joints.	45
4.3	Scree plot showing <i>proportions of variance</i> and <i>cumulative proportion of variance</i> explained by each component extracted using PCA.	53
4.4	Factor loadings of the features across the two extracted factors. The numeric values of the loadings are displayed within braces with the legend.	54
4.5	Receiver operating characteristic (ROC) plots for classifier models using different feature sets.	60
5.1	Information sufficiency across scenario groups and filters, in terms of mean values and standard error.	75
5.2	Photo satisfaction across scenario groups and filters, in terms of mean values and standard error.	77
5.3	Trade-off between protecting against information leaks and information sufficiency across filters, in terms of recognition accuracy (x-axis) and mean information sufficiency (y-axis). Note that <i>Silhouette</i> was not studied for any property related to <i>Environment</i>	78
5.4	Trade-off between protecting against information leaks and aesthetics across filters, in terms of recognition accuracy (x-axis) and mean visual aesthetics (y-axis). Note that <i>Silhouette</i> was not studied for any property related to <i>Environment</i>	80
6.1	An example illustrating how obfuscation and beautification change the utility aspects of an image: (a) an image without any alteration, (b) the image after a pixel obfuscation, and (c) the image after a pixel obfuscation applied to the food plate and a cartoon beautification on the other parts of the image.	89

6.2	Initial path model.	92
6.3	Example path model diagrams.	94
7.1	Histogram of variances for valence scores per photo.	122
7.2	Sum of squared distances of data samples to their closest cluster center for different number of clusters. The number of clusters was set to three based on the elbow method [75].	128
7.3	Mean (with 95% CI) sharing likelihood by humor type and photo-valence in the <i>baseline</i> condition.	130
7.4	Means (and 95% CI) of sharing likelihood	134
7.5	Mean likelihood (with 95% CI) to share photos by Females and Males across valence levels.	135
A.1	Receiver operating characteristic (ROC) plots for classifiers trained and tested on images with (a) 67% agreement and (b) 100% agreement among the survey participants.	186
A.2	Images used for attention check questions.	186
B.1	Screenshot of the application we developed to select filter levels.	190

LIST OF TABLES

4.1	Most frequent reasons found in the pilot study for classifying a person as a <i>Subject</i> and how many times each of them was selected in the main study.	47
4.2	Most frequent reasons found in the pilot study for classifying a person as a <i>Bystander</i> and how many times each of them was selected in the main study.	48
4.3	Correlation coefficients and effect sizes between the visual features and the reasons for classifying a person as a <i>subject</i> . All coefficients and effect-sizes are significant at $p < .001$ level.	48
4.4	Correlation coefficients and effect sizes between the visual features and the reasons for classifying a person as a <i>bystander</i> . All coefficients and effect-sizes are significant at $p < .001$ level.	50
4.5	Effectiveness of the selected features to classify ‘subject’ and ‘bystander’. The columns show odds-ratios and their 95% confidence intervals for each feature. All $p < 0.0001$	53
4.6	Mean and standard deviation of accuracy for classification using different feature sets across 10-fold cross validation.	56
5.1	Results of applying different filters to obscure food.	64
5.2	Scenarios and the recognition questions used in the survey.	65
5.3	Recognition accuracy for different filters across different scenarios. Recognition accuracies are shown as percentages, while subscripts and colors indicate whether each filter is effective (H), somewhat effective (M), or not effective (N) in preventing recognition, and asterices indicate significance: * is significant at $p < .05$, ** is significant at $p < .001$, and *** is significant at $p < 0.001$, after Bonferroni correction.	86

5.4 Privacy and utility trade-offs. For each filter, a green checkmark or red cross indicates whether that filter 1) protects privacy (i.e. recognition accuracy < 50% and odds-ratio < 0.05) (P), 2) provides sufficient information (I), 3) creates a satisfactory image (S), and 4) creates a visually appealing image (V).	87
6.1 Obfuscations and transformations used in this study. Each obfuscation was combined with each transformation, resulting in nine conditions. In addition, we included 3 obfuscation-only conditions, as well as a condition with the original, unaltered image, totaling 13 experimental conditions.	95
6.2 The six attributes and corresponding detection questions used in the survey.	96
6.3 Results of applying different obfuscations and beautifications.	98
6.4 Means and standard deviations of information content scores for different attributes.	104
6.5 Means and standard deviations of visual aesthetics scores for different attributes.	106
6.6 Means and standard deviations of photo satisfaction scores for different attributes.	107
7.1 Questions Presented for Each Condition	123
7.2 Z-scores of the cluster means along the four dimensions of humor.	128
7.3 Type II ANOVA Table (with Satterthwaite's method). (* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$). The effect size η_p^2 (partial η^2) can be interpreted as small if $\eta_p^2 = 0.01$, medium if $\eta_p^2 = 0.06$, and large if $\eta_p^2 = 0.14$ [117].	143
7.4 Correlation between the average meme-sharing likelihood of a participant and their past activities of sharing embarrassing or privacy-violating photos of themselves or others on social media (* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$).	144
A.1 Effectiveness of visual features used individually as predictors to classify <i>subject</i> and <i>bystander</i> . All χ^2 statistics are significant at $p < 0.0001$ level.	183

A.2 Correlation coefficients (r) between pairs of visual features. Each coefficient is significant at $p < .001$ level.	183
A.3 Variance inflation factor (VIF) of predictor variables when all predictors were used (Initial VIF) and after <i>Awareness</i> was removed (Updated VIF).	184
A.4 Results of predicting <i>high-level concepts</i> using image data. Columns show means and standard deviations of <i>loss</i> , <i>mean absolute error (MAE)</i> , and <i>mean squared error (MSE)</i> of a 10-fold cross-validation.	185
A.5 Percentage of participants agreed with the final classification label and number of photos with that agreement values.	185

CHAPTER 1

Introduction

Sharing photos on online social media platforms provides a natural mechanism for people to express themselves and interact with one another [104]. It plays an important role in creating and maintaining social connections in the online space [151]. Image macros, i.e., memes created using images, help to promote ideas and partially form the digital identity or online persona one desires [56]. Thus, the affordances provided by social media sites and applications such as Flickr [4], Instagram [5], Snapchat [7], Facebook [3], and WhatsApp [8] have caused a dramatic increase in the number of photos shared—by one estimate, more than 1.8 billion photos are posted to popular social media services each day [108]. This number is only expected to grow as traditional social media sites (e.g., Facebook) are outpaced in popularity by applications that focus more on visual means of communication (e.g., Instagram) [79, 129].

Despite offering several social benefits, photo-sharing activities on social media also pose threats to the privacy, security, and safety of people appearing in photos shared online. Photos are a rich source of visual data; in addition to the main subject matter, sometimes they contain incidental information that may be privacy-sensitive [12, 14, 41, 77, 95, 118, 189] and social media users have expressed regrets after posting photos that contain such information [167]. Further, once uploaded, photos (and associated metadata) reside on the cloud for an indefinite period of time. They can be re-shared with a larger (and sometimes unintended) audience or even publicly by the viewers. Thus, photos shared on social media can reach people outside of the ‘imagined audience’ [10, 120] and may lead to ‘context collapse’ [29, 30, 150]. To avoid privacy-violating incidents, social media users exercise self-censorship (e.g., posting only carefully selected photos) or even withdraw themselves completely from social media [190, 191, 216], which prevents them from receiving the social benefits that online platforms offer [81, 112, 206, 226].

Several technical and social mechanisms have been devised to help photo-sharers reduce privacy risks. One such approach is to limit access to the shared photos. Social media platforms (such as Facebook) offer access control mechanisms that enable one to share an item with a specific group of people. To ease this process at the individual image level, researchers have proposed algorithms that automatically suggest appropriate privacy settings while uploading a photo [193]. Other researchers have attempted to automatically detect privacy-sensitive content in images (such as certain places [201] and electronic screens [110]) or estimating an overall ‘privacy-sensitivity’ rating for an image [156]; such information can be conveyed to the photo-sharers to help them make an informed decision regarding the intended audience. In the case of multiple owners of an image with differing sharing and privacy preferences, researchers have proposed ways to allow the co-owners to negotiate and make sharing decisions in a collaborative fashion [197]. Social media users also exercise control offline to avoid sharing co-owned photos with undesired audiences [168, 169]. All these mechanisms are applicable only when people appearing in a photo can participate in the sharing decision. But, how can we protect the privacy of people who cannot exert any control over how their images are shared online? Reducing their privacy risks is the primary focus of this dissertation.

Photos taken in public places often contain ‘bystanders’— people who were captured just because of their physical co-location with the photo subjects. When these photos are shared online they reveal the identity, location, and other sensitive information about the bystanders. Researchers have proposed technical means to protect bystanders’ privacy [13, 184, 225], but they may not be scalable, since the processes involve manual steps such as requiring the bystanders to use specialized devices and applications. Moreover, such solutions impose an undue burden on the victims of privacy violations, where the bystanders have to be pro-active and vigilant to protect their privacy. These approaches also require the bystanders to broadcast personal data, such as facial features and locations, so that photographers can identify and remove them from their photos. These data

themselves are privacy-sensitive and sharing them expands the attack surface of privacy violations. To fill the gap of an automated and scalable solution to this problem, we have proposed a machine learning based model that can distinguish *bystanders* from *photo subjects* with high accuracy based on only the visual data in an image. It has the potential to protect bystanders' privacy at scale, as it can be incorporated easily into online social media and other cloud-based photo-hosting platforms.

Once bystanders in a photo are identified, their visual data can be kept private by, e.g., obscuring image regions containing them. This step can also be performed automatically using image filters such as blurring, pixelation, and masking. Despite such obfuscations having been in use for quite a long time, prior research investigating their effectiveness in obscuring sensitive information in images has largely focused on human identity [24, 77, 111, 125, 146]; although many other objects (e.g., computer monitors) and attributes (e.g., embarrassing activities) are also considered privacy-sensitive [12, 14, 41, 95, 118, 189]. This problem of accidentally sharing sensitive information with undesired audience is more frequent in the case of camera-based assistive devices for people with visual impairments [15]. However, systematic studies that investigate the effectiveness of image filters in obscuring sensitive image content at the object and attribute levels are lacking. Furthermore, online image-sharing is partly motivated by managing impressions [72] and eliciting social interactions from other people [206], so privacy-enhancing filters would require the obfuscated images to conserve enough viewer utility that the photo-sharers would still adopt them widely. But these aspects of commonly used filters (such as blurring and pixelation) have not been adequately studied in the prior literature. This dissertation assesses several popular filters in terms of both their ability to protect the privacy and preserve utility (e.g., whether the filtered images are visually pleasing and satisfactory to the viewers). Additionally, it sheds light on how the filters and three relevant utility variables (information content, visual aesthetics, and viewers' satisfaction) interact among themselves. Based on empirically validated interaction models, it offers guidance for designing better image obfuscations. Finally, findings from a user study that evaluated novel

image filters developed following that guideline are also presented.

Another way that people’s privacy may be violated is by creating memes using their photos and posting them on social media. Such memes are usually made using humorous or embarrassing photos found on the internet, possibly with further alterations such as adding a caption that is often demeaning to the photo subjects. They are then widely circulated and reach a large audience. Many times people appearing in these memes have experienced severe consequences in their professional and personal lives after memes featuring them went ‘viral’ [2, 21]. It is difficult to alleviate this problem by using only technical means. Therefore, this dissertation also explores a socio-technical approach to tackle this challenge. Concretely, we designed and evaluated *behavioral interventions* to nudge social media users to consider the implications of their meme-sharing acts for other people’s privacy and elicit privacy-respecting and pro-social behaviors. Furthermore, we identified several demographic and personal factors that influence photo-sharing decisions that would help in developing directed and personalized interventions.

Contributions This dissertation makes the following contributions.

Automatic detection of ‘bystanders’ in images. We have proposed a machine learning-based model to automatically detect bystanders in images so that their privacy can be protected by, e.g., obfuscating them using image filters. This model was trained on images from the *Google Open Image Dataset* [113] and tested on images from the *MS COCO Dataset* [126] with an overall detection accuracy of more than 84% [83]. It has the potential to protect people’s privacy at scale, as it can be easily integrated into social media platforms and provided as a cloud service for mobile-based photo-sharing applications.

Evaluating the effectiveness (to reduce privacy risks) and usability of image filters.

In an online study, we evaluated five commonly used image filters (e.g., blurring, masking, and pixelating) to assess how effectively they obscured 20 privacy-sensitive scene elements

and whether the filtered images preserved ‘sufficient’ utility for the viewers [84]. Findings from this study revealed that many of these filters, especially at a lower strength (e.g., a Gaussian blur filter with a small radius), failed to provide adequate privacy protection, despite being widely used. Further, a large portion of the ‘effective’ filters degraded the utility of the images to a degree that was unacceptable to the viewers.

Designing novel image filters. We have proposed a principled approach to designing new image filters to balance privacy and utility by modeling the interactions among privacy and utility variables (*information content*, *visual aesthetics*, and *viewers’ satisfaction*) and empirically validating the models. Following this approach, we have designed and evaluated new obfuscations by combining image filters and artistic image transforms [85].

Understanding social media users’ photo-sharing behaviors. We identify associations of behaviors related to sharing privacy-sensitive photos of other people with demographic factors (such as age and gender) and personality traits (e.g., how one uses humor to entertain oneself or advance social relationships). Additionally, we document how people with different ‘humor types’ reacted to interventions designed to encourage privacy-protective behaviors and provides evidence that generic interventions may not achieve the desired outcome. These findings add to the knowledge of understanding human behaviors in the context of online photo sharing and offer a guide to designing new interventions to encourage privacy-respecting and pro-social behaviors.

Dissertation Outline Chapter 2 of this document presents the problem statement and the thesis statement. Chapter 3 presents the relevant prior work in this field, including research about identifying privacy concerns among social media users and social and technical solutions to reduce privacy risks in the context of sharing images. In Chapter 4, we present details on the proposed machine learning-based solution to protect people’s privacy when they were captured in other

people's images as 'bystanders.' Chapter 5 presents the online study that evaluated five commonly used image filters in terms of their efficacy and acceptability to the images' viewers. Chapter 6 of this document details how we empirically validated the interactions among privacy protection and utility variables. It also provides findings from a study where we proposed and evaluated novel obfuscations. In Chapter 7, we present a series of experiments we undertook to understand people's photo-sharing behaviors, factors influencing those behaviors, and the effectiveness of behavioral interventions in reducing the sharing of photos that may violate people's privacy. In Chapter 9, we conclude with a summary of our findings and observations for future work. Finally, the Appendix includes additional findings and survey questionnaires.

CHAPTER 2

Problem and Thesis Statement

Photos capture memorable life-events, and sharing them with friends and family provides a natural mechanism for people to express themselves and interact with one another [104]. Participating in online social networks in general, and sharing photos in particular, offer various social benefits to people, including gaining gratification [81, 112, 226], making new and strengthening existing social relationships [206], and accumulating social capital [196]. The tremendous popularization of online social networks (OSNs) in the past decade, coupled with the ubiquity of image capturing devices such as traditional cameras, smartphones, and life-logging (wearable) devices, has resulted in a dramatic increase in photo capturing and sharing every day [57, 160]. As of June 2019, over 350 million images are uploaded each day to Facebook alone [192]. The volume of uploaded photos is expected to only rise as photo-sharing platforms such as Instagram and Snapchat continue to grow [34, 101]. A significant number of shared photos are taken in public places, thanks to the affordability of portable devices with cameras. Such photos often include ‘bystanders’— people who were captured just because of their physical co-location with the photo subjects and sharing these photos may violate the privacy of these bystanders.

In addition to their own photos, social media users re-share photos that they find on the internet. Recently, there has been unprecedented growth in publicly posting photo-based memes, which are made of photos with texts overlaid on them. Many memes depict embarrassing moments the subjects experienced, which is magnified using the accompanying texts. Memes are derived from not only photos featuring public figures (e.g., celebrities and politicians), but also from photos of the general public. These people can be maligned or embarrassed in front of a large population, leading to psychological distress and disruption in their professional and personal lives [2, 180].

Problem Statement: *The availability of photo-taking devices and the affordances provided by online social media have resulted in billions of photos being shared every day with an unbounded number of people. In many instances, social media users share photos that, in addition to themselves, include bystanders, who are usually strangers to the photo-owners/photographers. In many other instances, the shared photos consist entirely of strangers, as in the case of memes. In both scenarios, the shared photos violate the privacy of those strangers. While prior research has addressed the privacy issues of photo-sharers, privacy threats to the bystanders and people appearing in memes have not received adequate attention.*

2.1 Thesis

Researchers have proposed various technical means to protect the privacy of the bystanders in images [13, 26, 184, 185, 225]. These solutions allow the bystanders to communicate their privacy preferences with the photographers, e.g., using a smartphone app that broadcasts privacy policies using Bluetooth. But this approach relies on the bystanders, who are the victims of privacy violations, to be proactive in keeping their data private. Further, this approach requires the bystanders and photographers to use specialized and compatible devices or applications. Most of the proposed solutions require the bystanders to share sensitive data (such as facial features [13, 225] and location [184]) so that the photographers can identify them and apply the intended privacy policy. Some of them necessitate broadcasting bystanders' privacy preferences publicly (e.g., using visual markers [26] or hand gestures [185]), which in itself might be a privacy violation. Finally, none of the proposed solutions apply to the photos that have been previously taken and/or stored in devices or the cloud. Thus, to reduce the privacy risks of the bystanders at scale, an automated mechanism is required that works independently of photo-taking devices and does not require the bystanders and photographers to use compatible applications, does not increase the risk of privacy violations by transmitting additional information, and can be applied to all past and future photos.

A machine learning-based model to automatically distinguish bystanders from photo subjects using only the visual data present in the images might be a feasible first step toward this goal.

Once the bystanders in a photo are identified, how can their privacy be ensured even if the photo is shared online? Limiting the dissemination of the photo (e.g., by using the privacy settings that are offered by social media platforms) is not an effective solution, since sharing it with a smaller number of people may still violate the bystanders' privacy. Moreover, this does not prevent someone in the first order audience from re-sharing the image with a larger group of people or publicly. Not sharing the image at all would guarantee privacy protection, but this is against one of the primary motivations of using social media and will deprive the users of many social benefits. A more viable solution may be to alter the regions in these images that contain bystanders to obscure privacy-sensitive information such as their identity, activity, and facial expressions. But few studies have systematically investigated how effective image filters/transforms (e.g., scrambling, blurring, and pixelating) are in properly obscuring such information. Another important concern in the social media context is how usable and acceptable such transforms are among the photo-sharers and viewers. Applying filters to an image would inevitably remove some information from it and may leave the filtered image visually unappealing. This reduction in utility may be unacceptable to the photo-sharers, whose primary motivations to share photos include conveying information and seeking acceptance, appreciation, and validation from peers [130, 151].

How do we prevent privacy violations resulting from the circulation of memes on social media? Technical approaches seem to be inadequate in combating this challenge. Rather, stimulating privacy-respecting and pro-social behaviors among the social media users— who are simultaneously the content producers, propagators, and consumers— may be a feasible approach toward solving this problem. To advance in this direction, it is crucial to understand people's decision-making processes in the context of sharing privacy-sensitive photos and the factors that affect those decisions. Such an understanding would help us to redesign social platforms with implicit cues to reinforce people's

sense of propriety and explicit priming to nudge them toward privacy-protective behaviors.

Thesis statement: *Recent advancements in machine learning and computer vision can be leveraged to develop automated, scalable, and usable solutions to protect the privacy of bystanders. A promising approach to prevent privacy violations through sharing memes is identifying the factors that affect the meme-sharing behaviors of social media users and developing behavioral interventions to discourage such activities.*

2.1.1 Dissertation Outline

The work in this thesis is divided into four chapters:

1. **Designing machine learning-based models to automatically classify ‘bystander’ and ‘subject’ in images:** Can subjects and bystanders in photos be distinguished by machine learning models using only visual data present in those photos, given that these two concepts are highly contextual? We attempt to address this question by learning how humans categorize subjects and bystanders in photos and building machine learning models following the reasonings our surveyed population used.
2. **Evaluating the privacy-protection capability and usability of image filters:** Besides identity, many other types of information about people (e.g., facial expressions) and other objects (e.g., text on an electronic screen) are considered privacy-sensitive. Such incidental information leaks may even threaten the privacy of the photo subjects (and not only the bystanders). In this chapter, we evaluate how well image filters can obscure such information while preserving utility for the image viewers.
3. **Designing novel image obfuscations:** Our evaluation of image filters revealed privacy-utility trade-offs: filters that effectively protected privacy also reduced the utility of the images (e.g., visual aesthetics). How can utility be improved without compromising privacy? Based

on path model analyses on experimental data, this chapter proposes novel image obfuscations to improve utility and presents the findings of a user study evaluating them.

4. **Individual differences and photo-sharing behaviors:** Technical solutions to detect and obfuscate privacy-sensitive visual data are less applicable in the context of sharing memes on social media. We explore an alternative approach to address this issue: understanding how people make meme-sharing decisions and designing interventions to alter their behavior. This chapter presents a study that discovered links between how people differ in *using humor* and their preference in sharing memes.

2.1.2 Designing Machine Learning-Based Models to Automatically Classify ‘Bystander’ and ‘Subject’ in Images

Chapter 4 details our approach to protecting bystanders’ privacy by automatically distinguishing them from photo subjects so that they can be, e.g., obfuscated to obscure their identity and other sensitive information. With annotated data collected through a user study, we built and evaluated several machine learning models to classify people in images as ‘bystanders’ or ‘subjects’. Our best performing model, which performs the classification task in two steps, achieved over 85% classification accuracy. First, it extracts features from several existing deep neural network models including ResNet50 [89] and OpenPose [36], and uses these features to infer higher-level features such as whether a person was *posing* for a photo. In the second step, it uses those inferred feature values to classify a person as a subject or a bystander in an image. The higher-level features were identified based on what ‘concepts’ our study participants used to distinguish between bystanders and subjects, and whether significant associations existed between those features and *why* the participants labeled a person in an image as a subject or a bystander. We compared this model with several other models; one of them was trained directly using the ‘lower’ level features, but this model outperformed all other models by a large margin. Importantly, the classification decisions

made by this model can be easily explained, as they are based on a few *high level* features that are related to the intuitive visual characteristics of a person in a photo that humans use to make such distinctions.

The following are the key contributions of this chapter.

1. Through a user study, we identify the rationales humans use to distinguish *bystanders* from *subjects*.
2. To train machine learning models to detect bystanders using image data, we proposed a set of intuitive features. We validated them using data from the study, and identified a subset of features that are *minimally correlated* among themselves to use in the training phase.
3. We trained a model using those features that yielded high classification accuracy in classifying, even when assigning these roles was not straightforward for human annotators.
4. As the features represent the humans' high-level, intuitive conceptualizations of 'bystander' and 'subject,' the decisions made by the model can be easily explained by examining this handful of features, which improves the system's transparency.

2.2 Evaluating the Privacy-Protection Capability and Usability of Image Filters

In Chapter 5, we assess the effectiveness (to protect privacy) and usability of five commonly used image filters (e.g., blurring, pixelating, and silhouette). The filters were applied to specific regions of images containing privacy-sensitive information (e.g., facial expressions, gender, and a person's ethnicity). These filtered images were shown to the participants of an online study. The ability of a filter to protect privacy was determined based on whether the participants could correctly recognize the obscured information (e.g., a person's gender). Each filter's usability was measured by the perceived *information content* (i.e., whether there was 'enough' information to understand the filtered image), whether the filtered image was visually pleasing, and whether the filtered image

was satisfactory to the participants as viewers. Analyses from the experimental data revealed that in most of the cases, the filters *failed* to properly obscure the intended information. In cases where the filters successfully protected privacy, the resulting images were not satisfactory to the participants as they reduced too much information and/or visual aesthetics.

The main contributions of this chapter are:

1. We applied five commonly used filters (with varying strength, totaling 11 transformations) over different objects (e.g., people, electronic screens, and paper documents) to obscure twenty properties (e.g., facial expression, activity, and text) that were identified as privacy sensitive in prior studies. The filters were assessed on how well they obscured the intended information.
2. These filters were also evaluated based on three utility variables: *information content*, *visual aesthetics*, and *viewers' satisfaction*.
3. To the best of our knowledge, this was the first study to assess image transforms at property or attribute level, rather than at the object level.

2.3 Designing Novel Image Obfuscations

In Chapter 6, we analyzed how the three utility variables (i.e., information content, visual aesthetics, and satisfaction) affect each other, and based on their interactions we proposed and evaluated novel image obfuscations. Path model analyses on the data from the previous study (described in Chapter 5) revealed that the *information content* of an image *positively* affects both *visual aesthetics* and its *viewers' satisfaction*. Further, *visual aesthetics* also *positively* impacts *viewers' satisfaction*. These imply that a filtered image can be made more satisfactory to the viewers by enhancing its information content and/or improving its visual aesthetics. To empirically validate these findings, we experimented with creating novel obfuscations by combining image filters (such as blurring) with artistic transforms (e.g., cartoonization). These new obfuscations were evaluated

in terms of the same three utility variables through a new study. We found that the artistic transforms enhanced the visual aesthetics of filtered images in some cases, but the improvement was not significant enough to recover from the loss in satisfaction caused by the filters.

The primary contributions of this chapter include:

1. We modeled the interactions among filters and three utility variables (information content, visual aesthetics, and viewers' satisfaction) through *path model analysis*. These models were empirically validated using experimental data.
2. Path model analyses revealed how the filters and the utility variables affect each other. In particular, it suggested a mechanism to improve *viewers' satisfaction*: by enhancing any or both of the *information content visual aesthetics*.
3. Following this mechanism, we applied artistic transforms to beautify the privacy-enhanced images. The resulting obfuscations were assessed in terms of the three utility variables through a new user study.

2.4 Individual Differences and Photo-sharing Behaviors

In Chapter 7, we detail an experiment where we attempted to understand the link between how people use *humor* and their preferences in sharing memes on social media. The second goal of this experiment was to determine whether people's reactions to privacy nudges would differ as a function of their 'humor type'. To answer these questions, we conducted an online study where participants viewed memes in one of three randomly assigned experimental conditions and indicated the likelihood of them sharing those memes on their social media accountst. The memes used in that study were rated according to their valence (i.e., how *positively* or negatively a meme portrayed the photo subjects) by 400 participants in a separate study. The three experimental conditions included one control condition (i.e., without any interventions) and two priming conditions where

participants were instructed to imagine themselves as the photo subjects or consider the privacy of the photo subjects. We also collected data about the participants' usage of humor using the *Humor Style Questionnaire* (HSQ) [133]. Participants were divided into three groups based on their 'humor type.' We found that 'humor endorsers'— participants who frequently use humor to entertain themselves or other people— were more likely to share memes that portrayed the subjects negatively. We also replicated the paradoxical findings reported by Amon *et al.* [16]: participants who were primed to consider photo subjects' privacy demonstrated a *higher* sharing likelihood compared to the control group. Our study provides further insights into how this seemingly paradoxical behavior relates to participants 'humor type'. We found that 'humor deniers' – participants who infrequently use humor to entertain themselves or others – intended to share *more* after the privacy nudge, but this behavior was not observed for participants in the other two humor categories.

The primary contributions of this chapter include:

1. We identified 'humor type' (i.e. how people use humor to entertain themselves or advance social relationships) as an important factor affecting photo-sharing decisions.
2. The study provides evidence that people's reactions to behavioral interventions may differ depending on their 'humor type'; thus, personalized interventions may yield better outcomes.
3. Based on extensive literature review on human psychology, this chapter offers insights on the likely reason behind participants' seemingly paradoxical behavior.

CHAPTER 3

Related Work

In this chapter, we present relevant prior work that has been done to understand people’s motivations for sharing photos on social media, how sharing photos may threaten the privacy of people who appear in them, and what technical and social measures have been proposed to mitigate privacy risks in this context.

3.1 Sharing Photos on Social Media

Extensive research has been conducted to understand why people share photos on online social media. Malik *et al.* applied Uses and Gratification theory to understand what drives the sharing of photos on Facebook and identified affection, attention seeking, disclosure, information sharing, habit, and social influence as the primary motivators [130]. Oeldorf-Hirsch and Sundar identified four classes of gratifications that inspire people to post photos online: seeking and showcasing experiences, technological affordances, social connection, and reaching out [151]. Their results suggest that people attempt to fulfill social needs through sharing photos in the virtual space [151]. Sung *et al.* studied social and psychological motives for selfie-posting behaviors on social networking sites and established four motivations: attention seeking, communication, archiving, and entertainment. Regarding image macros or memes, they are shared to communicate profound philosophical ideas or contemporary political issues, as well as just for entertainment [139].

3.2 Privacy Risks in the Context of Sharing Photos on Social Media

Prior research has identified what information people consider to be threatening to their privacy when revealed to others through photographs [12, 14, 41, 77, 95, 110, 118, 165, 189, 201]. The most frequently mentioned privacy-sensitive attribute that photos may reveal is identity. Presumably, it

may matter less in the social media context when the photos reveal the photo-sharers' identity to their online contacts, while it might be the most sensitive information to reveal from the bystanders' perspective. Other information about people that is considered privacy-sensitive include facial expression, gender, race, and activities [14, 41, 118, 141, 188, 189]. Prior research has also identified people's concerns regarding leaking private information through other objects (e.g., the content of computer screens [110] and text on paper [136]), places [118], and living conditions (e.g., a messy room [41]). Metadata associated with images also leak sensitive information that is particularly threatening for the bystanders. For example, image metadata and/or accompanying text (e.g., 'status update' in Facebook) may contain location and date information, which can be combined with facial identity to learn (manually or automatically) people's location at a particular time. This might be desirable for the photo-sharers (e.g., while sharing photographs of a summer vacation or eating out with friends), but very concerning for the bystanders.

Social media users do not only post photos of themselves or that they had taken, but also re-share photos posted by others or found on the internet, possibly with additional alterations (e.g., adding text to make a 'meme') to fit specific purposes. Often photos in which subjects were portrayed in embarrassing ways, are selected to make memes, and the embarrassment is highlighted or amplified with additional texts. This severely undermines the photo subjects' privacy, as well as social and professional impression [16, 49].

Beyond privacy threats to individuals, at a collective level, the abundance of publicly available photos aid in building and deploying automated tools to identify and track people online [11, 181, 194]. Such technologies are already being used by law enforcement agencies to find suspects [17, 91, 134] and can easily be abused for surveillance, targeted advertising, and stalking, which threaten people's privacy, autonomy, and even physical safety.

3.3 Protecting Bystanders' Privacy

Prior work on alleviating bystanders' privacy risks can be broadly divided into two categories—techniques to handle images i) stored in the photo-capturing device and ii) after being uploaded to the cloud (Perez et al. provide a taxonomy of proposed solutions to protect bystanders' privacy [159]).

3.3.1 Privacy Protection in the Moment of Photo Capture

3.3.1.1 Preventing image capture

Various methods have been proposed to prevent capturing photographs to protect the privacy of nearby people. One such method is to temporarily disable photo-capturing devices using specific commands that are communicated by fixed devices (such as access points) using Bluetooth and/or infrared light-based protocols [203]. One limitation of this method is that the photographers would have to have compliant devices. To overcome this limitation, Truong *et al.* proposed a ‘capture resistant environment’ [208] consisting of two components: a camera detector that locates camera lenses with charged coupled devices (CCD) and a camera neutralizer that directs a localized beam of light to obstruct its view of the scene. This solution is, however, effective only for cameras using CCD sensors. A common drawback shared by these location-based techniques [203, 208] is that it might not be feasible to install them in every location.

Aditya et al. proposed I-Pic [13], a privacy enhanced software platform where people can specify their privacy policies regarding photo-taking (i.e., whether someone is allowed to take photos or not), and compliant cameras can apply these policies over encrypted image features. Although this approach needs the bystanders to participate actively, Steil *et al.* proposed PrivacEye [195], a prototype system to automatically detect and prevent capturing images of people by automatically covering the camera with a shutter. Although the bystanders do not take any action to protect their privacy, PrivacEye [195] considers every person appearing in an image, limiting its applicability in

more general photography settings.

The main drawback with these approaches is that they seek to completely prevent the image from being captured. In many cases, this may be a heavy-handed approach where removing or obscuring bystanders is more desirable.

3.3.1.2 Obscuring bystanders

Several proposed solutions to protect bystanders' privacy utilize image-obfuscation techniques to obscure bystanders in images, instead of preventing image capture in the first place. Zhang *et al.* developed COIN [225], which lets its users broadcast privacy policies and identifying information in much the same way as I-Pic [13] and obscure identified bystanders. In the context of wearable devices, Dimiccoli *et al.* developed deep-learning-based algorithms to recognize the activities of people in egocentric images degraded in quality to protect the bystanders' privacy [55].

Another set of proposed solutions enables people to specify privacy preferences *in situ*. Li *et al.* present PrivacyCamera [121], a mobile application that handles photos containing at most two people (either one bystander, or one target and one bystander). Upon detecting a face, the app sends notifications to nearby bystanders who are registered users of the application using short-range wireless communication. The bystanders respond with their GPS coordinates, and the app then decides if a given bystander is in the photo based on the camera's position and orientation. Once the bystander is identified (e.g., the smaller of the two faces), their face is blurred. Ra *et al.* proposed Do Not Capture (DNC) [164], which tries to protect bystanders' privacy in more general situations. Bystanders broadcast their facial features using a short-range radio interface. When a photo is taken, the application computes the motion trajectories of the people in the photo, and this information is then combined with facial features to identify bystanders, whose faces are then blurred.

Several other papers outline techniques that allow users to specify default privacy policies that

can be updated based on context using gestures or visual markers. Using Cardea [184], users can state default privacy preferences depending on the location, time, and presence of other users. These static policies can be updated dynamically using hand gestures, giving users the flexibility to tune their preferences depending on the context. In a later work, Shu *et al.* proposed an interactive visual privacy system that uses tags instead of facial features to obtain a given user’s privacy preferences [185]. This is an improvement over Cardea’s system since facial features are no longer required to be uploaded. Instead, different graphical tags (such as a logo or a template, printed or stuck on clothes) are used to broadcast privacy preferences, where each of the privacy tags refers to a specific privacy policy, such as ‘blur my face’ or ‘remove my body.’

In addition to the unique limitations of each of the aforementioned techniques, they also share several common drawbacks. For example, solutions that require transmitting bystanders’ identifying features and/or privacy policies over wireless connections are prone to Denial of Service attacks if an adversary broadcasts this data at a high rate. Further, there might not be enough time to exchange this information when the bystander (or the photographer) is moving and goes outside of the communication range. Location-based notification systems might have limited functionality in indoor spaces. Finally, requiring extra sensors, such as GPS for location and Bluetooth for communication, may prevent some devices (such as traditional cameras) from adopting them.

3.3.2 Protecting Bystanders’ Privacy in Images in the Cloud

Another set of proposed solutions attempts to reduce bystanders’ privacy risks after their photos have been uploaded to the cloud. Henne *et al.* proposed SnapMe [90], which consists of two modules: a client where users register, and a cloud-based watchdog that is implemented in the cloud (e.g., online social network servers). Registered users can mark locations as private, and any photo taken in such a location (as inferred from image metadata) triggers a warning to all registered users who marked it as private. Users can additionally let the system track their locations and send

warning messages when a photo is captured nearby their current location. The users of this system have to make a privacy trade-off, since increasing visual privacy will result in a reduction in location privacy.

Bo *et al.* proposed a privacy-tag (a QR code) and an accompanying privacy-preserving image sharing protocol [26] which could be implemented on photo sharing platforms. The preferences from the tag contain a policy stating whether or not photos containing the wearer can be shared, and if so, with whom (i.e. in which domains/PSPs). If sharing is not permitted, then the privacy tag wearer’s face is replaced by a random pattern generated using a public key from the tag. Users can control dissemination by selectively distributing their private keys to other people and/or systems to decrypt the obfuscated regions. More recently, Li and colleagues proposed HideMe [122], a plugin for social networking websites that can be used to specify privacy policies. It blurs people who indicated in their policies that they do not want to appear in other peoples’ photos. The policies can be specified based on scenario instead of for each image.

A major drawback of these cloud-based solutions is that the server can be overwhelmed by uploading a large number of fake facial images or features. Even worse, an adversary can use someone else’s portrait or facial features and specify an undesirable privacy policy. Another limitation is that they do not provide privacy protection for the images that were uploaded in the past and are still stored in the cloud.

3.4 Approaches to Reducing Privacy Risks

3.4.1 Limiting Dissemination

Online platforms have implemented mechanisms to limit the dissemination of a shared item among the intended audience. For example, Facebook allows users to specify who can view a shared photo. To ease this process at the individual image level, researchers have proposed algorithms that automatically suggest appropriate privacy settings while uploading a photo [193]. But such mechanisms

cannot prevent one from re-sharing an item with a larger audience on the same or different platforms. In the case of multiple people co-owning a photo (e.g., a group photo), researchers have invented mechanisms to enable negotiations among the co-owners who may have different privacy preferences [197]. Yasmeen *et al.* reported how undergraduate students negotiate offline to limit the dissemination of photos taken by others that depict them in an unflattering way, as it would adversely affect their reputation and future employment prospects [168]. Such solutions, however, are not applicable while sharing photos of strangers (e.g., memes), as they cannot exercise ownership even though they are the subjects in the photos being shared.

3.4.2 Reducing Privacy Risks by Obfuscating Sensitive Scene Elements

3.4.2.1 Obfuscation mechanisms

One way to prevent disclosing private information in an image is to only allow the people in the desired audience group to view the full image and ‘hide’ the private region of the image from the rest. PuPPIeS [88] and P3 [165] follow this approach; users can encrypt parts of images that contain sensitive information before sharing them via social networking sites or storing them in the cloud. POP [224] supports masking and blurring in addition to encrypting sensitive image regions before uploading to cloud servers. These approaches focus on encryption techniques to cryptographically ‘lock’ sensitive regions of images that can be ‘unlocked’ only by authorized users. But due to the volume of images shared on social media, it is not feasible to specify who are the desired audience for every image shared, and many images are meant for public consumption. Further, revealing incidental information, such as personal belongings, affiliations, and computer monitor contents even to the desired audience may pose risks to the photo-sharers’ privacy [95, 211]. Obfuscating such content using image filters may be more effective in protecting privacy in such situations.

Blurring and pixelation are two of the most commonly used filters in existing research and applications (e.g., [25, 100]). YouTube blurs faces in videos [25], while Google Street View obscures

faces and license plates to avoid identity leakage [63]. In the context of remote collaboration via live video feed, Boyle *et al.* [31] studied how blurring and pixelating affect privacy and awareness, and Hudson *et al.* proposed techniques such as representing people’s movement using dark pixels overlaid on a static image to reduce privacy risks while keeping information required for the collaboration [97]. Li *et al.* studied the effectiveness of several filters, including silhouette, point-light, and in-painting, to protect people’s identity. Other studies have investigated different forms of face de-identification [24, 77, 111, 146] for privacy protection. Hassan *et al.* [86] employed ‘cartooning’ transforms on videos in order to conceal sensitive scene elements but still convey certain characteristics.

Recently, researchers have focused their attention on obfuscating at the ‘attribute’ level, rather than at the object level. That is, they attempt to obscure specific attributes (e.g., facial expressions) of an object while revealing other attributes (e.g., identity). Sim *et al.* provided a mechanism to control which aspect (e.g., identity) of a face to obscure while retaining others (e.g., gender, race) as recognizable [188]. Mirjalili and Ross designed a technique that modifies a face image in a way that allows for automatic face recognition while preventing gender classifiers from inferring the gender from the face [141]. Ren *et al.* developed a system using adversarial machine learning to remove faces from video frames while retaining enough information for automatic action recognition [173].

3.4.2.2 Privacy-Utility Trade-Offs of Image Filters

Evidence from the literature suggests that many of the commonly used filters fail to prevent humans and/or machine learning models to recognize obfuscated contents. Gross *et al.* [76] demonstrated that for human faces, blurring and pixelating often either do not obscure enough details to provide adequate privacy or obscure so much that they destroy the utility of the video. The work of Brkic *et al.* [32] have shown that some obfuscation techniques can be defeated by neural network-based attacks. McPherson *et al.* [136] used deep learning algorithms to correctly identify faces and

recognize objects and handwritten digits even after they were blurred, pixelated, or encrypted with P3 [165].

With regard to image filters' utility, Li *et al.* reported that many of the filters (e.g., masking, bar, and point-light) that were effective in protecting privacy were not satisfactory to the viewers because they lowered utility variables such as information content [125]. These studies highlight the privacy-utility trade-offs. Filters that fail to strike the balance between the two are unlikely to be adopted by the photo sharers, since they tend to upload aesthetically-pleasing photos on social networks in order to manage the impression that they leave on others [186] and cultivate social interactions such as re-sharing by the viewers [104].

To improve photos' aesthetic utility, many studies have been conducted in the image processing and computer vision fields. Recent work in deep learning-based image style transfer has created beautifications that mimic the style of particular artists [68, 223]. Other systems also attempt to generate images with specific artistic effects like cartooning [87, 114] or water-coloring [28]. We draw on several of these beautification techniques, studying them in the context of how they affect perceived visual aesthetics and viewer satisfaction for obfuscated images.

3.5 Employing Behavioral Interventions to Reduce Privacy Risks

Behavioral interventions, such as ‘privacy nudge,’ have been employed to help people make ‘better’ decisions regarding privacy and security in many contexts (see [9] for a review). But so far, researchers had limited success in altering people’s behavior through interventions; one reason for this might be that the interventions were generic (i.e., they were developed for the ‘average’ person) and designed without considering individual differences [109]. Recently, Amon *et al.* reported a study where they designed behavioral interventions to discourage the sharing of photos that may violate the photo subjects’ privacy and evaluated the intervention in an online setting [16]. They reported surprising findings from that study—participants who were primed to consider the pri-

vacy implications of their photo-sharing acts demonstrated a *higher* sharing likelihood compared to the control group [16]. This result underscores the necessity to understand the demographic factors, personality traits, and contextual factors that influence photo-sharing behaviors and privacy concerns and design more directed and personalized interventions based on this knowledge. Prior research has identified many individual differences relevant to online photo-sharing contexts. Many researchers also designed and evaluated personalized interventions in several privacy/security related situations. But research on mitigating people’s privacy risks when they are featured in memes has been almost non-existent. Below, we review the existing literature in this direction.

3.5.1 Effects of Personality Traits on Photo-sharing Behaviors

Numerous studies have looked into individual differences in personality traits and their effects on social media usage and photo-sharing activities [45, 98, 143, 176]. Multiple studies found that extraversion [45, 98] and openness to new experiences [45] were positively correlated with social media usage and photo-sharing frequency, while emotional stability negatively predicted both of them [45, 98]. Ryan and Xenos reported that Facebook usage was positively correlated with narcissism and loneliness, while negatively correlated with conscientiousness [176]. Moore and McElroy documented that people high in conscientiousness made significantly fewer posts on Facebook about themselves and others [143]. Additionally, more conscientious people expressed more regret about posts with inappropriate content than did less conscientious people; this was also true for agreeableness [143]. Regarding sharing photos of strangers, Amon *et al.* administered a shorter version [166] of the five-factor model of personality traits, but did not find any association between the personality traits and meme-sharing likelihoods.

3.5.2 Demographic Differences in Photo-Sharing Behaviors

A sizable amount of prior literature was dedicated to understand how people’s social media usage, information-disclosing behaviors in general and photo-sharing habits in particular, and associated

privacy concerns vary depending on demographic factors including age and gender. Several studies found that women spend more time on social media platforms [94, 143] than men, even though women were also more concerned about their privacy [94, 182, 202]. Women were also identified to be more risk-averse [35, 39] and likely to take privacy-protective measures, such as activating privacy settings and un-tagging themselves from posts they did not want to be associated with [202], compared to men. Older people were found to be more concerned about privacy risks [222, 228] and they proactively protect their data compared to younger adults [222]. Regarding education, the findings have been mixed— Zukowski and Brown reported that internet users with higher levels of education were less concerned about information privacy than internet users with lower levels of education [228]; but Sheehan reported the opposite findings [183]. Regarding sharing photographs in social media, several studies found that women post more photos than men [94, 138, 143]. Biolcati and Passini documented gender differences in selfie posting behaviors – women posted more group selfies than men did, but no difference was found for own selfies. Prior research is almost non-existent with respect to posting photos of strangers, except the work of Amon *et al.* , who reported that female participants were significantly less likely to share strangers' photos than male participants, unless the photo represented the subjects *very positively* [16].

3.5.3 Personalized Interventions

Wisniewski *et al.* reported that social media users differ in how they manage their privacy and argued that behavioral interventions may be seen as hindrances if they are not aligned with the users' established privacy behaviors [217]. Based on data about how users engaged in protecting their privacy, the authors empirically established six 'privacy profiles' and recommended to design personalized nudges based on these profiles to elicit privacy-protective behaviors in the context of disclosing users' own data on social media [217]. Misra and Such developed a personal agent using users' profile information, context, and network structure to help them decide whom to share

information with [142]. In the context of IoT (Internet of Things) privacy, Bahirat *et al.* learned information-disclosing behaviors of IoT users and created privacy settings based on the frequently observed disclosing preferences [19]. These settings were recommended to new users as defaults, which were preferred to naive default settings by their study participants [19]. Researchers have also put these ideas into practice. For example, Liu *et al.* implemented a personalized app permission assistant that was well accepted by the study participants [127].

CHAPTER 4

Machine Learning Based Models for Automatic Classification of ‘Bystander’ and ‘Subject’ in Photos

In this chapter, we present our proposed machine learning based model that was trained to automatically distinguish ‘bystanders’ from photo-subjects. This work was done in collaboration with Mario Fritz, David Crandall, and Apu Kapadia and published as “Automatically Detecting Bystanders in Photos to Reduce Privacy Risks” in the IEEE Symposium on Security and Privacy, 2020 [83].

4.1 Introduction

As described in Chapter 3, the ubiquity of image capturing devices (such as traditional cameras, smartphones, and life-logging (wearable) cameras) and tremendous popularity of social media platforms have resulted in creating and sharing billions of photos everyday. Since a significant portion of shared photos are taken in public places, often they contain ‘bystanders’ – people who were photographed incidentally without actively participating in the photo shoot. Such incidental appearances in others’ photos can violate the privacy of bystanders, especially since these images may reside in cloud servers indefinitely and be viewed and (re-)shared by a large number of people. This privacy problem is exacerbated by computer vision and machine learning technologies that can automatically recognize people, places, and objects, thus making it possible to search for specific people in vast image collections [11, 181, 194]. Indeed, scholars and privacy activists called it the ‘end of privacy’ when it came to light that Clearview – a facial recognition app trained with billions of images scraped from millions of websites that can find people with unprecedented accuracy and speed – was being used by law enforcement agencies to find suspects [17, 91, 134]. Such capabilities can easily be abused for surveillance, targeted advertising, and stalking that threaten peoples’

privacy, autonomy, and even physical security.

Recent research has revealed peoples' concerns about their privacy and autonomy when they are captured in others' photos [52, 145, 168]. Conflicts may arise when people have different privacy expectations in the context of sharing photographs in social media [198], and social sanctioning may be applied when individuals violate collective social norms regarding privacy expectations [67, 170]. On the other hand, people sharing photos may indeed be concerned about the privacy of bystanders. Indeed, some photographers and users of life-logging devices report that they delete photos that contain bystanders [14, 95], e.g., out of a sense of "propriety" [95].

A variety of measures have been explored to address privacy concerns in the context of cameras and bystanders. Google Glass's introduction sparked investigations around the world, including by the U.S. Congressional Bi-Partisan Privacy Caucus and Data Protection Commissioners from multiple countries, concerning its risks to privacy, especially regarding its impact on non-users (i.e., bystanders) [58, 152]. Some jurisdictions have banned cameras in certain spaces to help protect privacy, but this heavy-handed approach impinges on the benefits of taking and sharing photos [81, 112, 206, 226]. Requiring that consent be obtained from all people captured in a photo is another solution but one that is infeasible in crowded places.

Technical solutions to capture and share images without infringing on other people's privacy have also been explored, typically by preventing pictures of bystanders from being taken or obfuscating parts of images containing them. For example, Google Street View [74] treats every person as a bystander and blurs their face, but this aggressive approach is not appropriate for consumer photographs since it would destroy the aesthetic and utility value of the photo [84, 155]. More sophisticated techniques selectively obscure people based on their privacy preferences [13, 164, 184, 185, 225], which are detected by nearby photo-taking devices (e.g., with a smartphone app that broadcasts preference using Bluetooth). Unfortunately, this approach requires the bystanders – the victims of privacy violations – to be proactive in keeping their visual data private. Some proposed solutions

require making privacy preferences public (e.g., using visual markers [26] or hand gestures [185]) and visible to everyone, which in itself might be a privacy violation. Finally, these tools are aimed at preventing privacy violations as they happen and cannot handle the billions of images already stored in devices or the cloud.

We explore a complementary technical approach: automatically detecting bystanders in images using computer vision. Our approach has the potential to enforce a privacy-by-default policy in which bystanders' privacy can be protected (e.g., by obscuring them) without requiring bystanders to be proactive and without obfuscating the people who were meant to play an important role in the photo (i.e., the subjects). It can also be applied to images that have already been taken. Of course, detecting bystanders using visual features alone is challenging because the difference between a subject and a bystander is often subtle and subjective, depending on the interactions among people appearing in a photo as well as the context and the environment in which the photo was taken. Even defining the concepts of 'subject' and 'bystander' is challenging, and we could not find any precise definition in the context of photography; the Merriam-Webster dictionary defines 'bystander' in only a general sense as "one who is present but not taking part in a situation or event: a chance spectator," leaving much open to context as well as social and cultural norms.

We approach this challenging problem by first conducting a user study to understand how people distinguish between subjects and bystanders in images. We found that humans label a person as 'subject' or 'bystander' based on social norms, prior experience, and context, in addition to the visual information available in the image (e.g., a person is a 'subject' because they were interacting with other subjects). To move forward in solving the problem of automatically classifying subjects and bystanders, we propose a set of high-level visual characteristics of people in images (e.g., willingness to be photographed) that intuitively appear to be relevant for the classification task and can be inferred from features extracted from images (e.g., facial expression [123]). Analyzing the data from this study, we provide empirical evidence that these visual characteristics are indeed

associated with the rationale people utilize in distinguishing between subjects and bystanders. Interestingly, exploratory factor analysis on this data revealed two underlying social constructs used in distinguishing bystanders from subjects, which we interpret as ‘visual appearance’ and ‘prominence’ of the person in a photo.

We then experimented with two different approaches for classifying bystanders and subjects. In the first approach, we trained classifiers with various features extracted from image data, such as body orientation [36] and facial expression [123]. In the second approach, we used the aforementioned features to first predict the high-level, intuitive visual characteristics and then trained a classifier on these estimated features. The average classification accuracy obtained from the first approach was 76%, whereas the second approach, based on high-level intuitive characteristics, yielded an accuracy of 85%. This improvement suggests that the high-level characteristics may contain information more pertinent to the classification of ‘subject’ and ‘bystander’, and with less noise compared to the lower-level features from which they were derived. These results justify our selection of these intuitive features, but more importantly, it yields an intuitively-explainable and entirely automatic classifier model where the parameters can be reasoned about in relation to the social constructs humans use to distinguish bystanders from subjects.

4.2 Study Method

We begin with an attempt to define the notions of ‘bystander’ and ‘subject’ specific to the context of images. According to general dictionary definitions,¹²³ a bystander is a person who is *present and observing* an event *without taking part* in it. But we found these definitions to be insufficient to cover all the cases that can emerge in photo-taking situations. For example, sometimes a bystander may not even be *aware of* being photographed and, hence, not observe the photo-taking event. Other times, a person may be the subject of a photo without actively participating (e.g., by posing) in the

¹<https://www.merriam-webster.com/dictionary/bystander>

²<https://dictionary.cambridge.org/us/dictionary/english/bystander>

³<https://www.urbandictionary.com/define.php?term=bystander>

event or even noticing being photographed, e.g., a performer on stage being photographed by the audience. Hence, our definitions of ‘subject’ and ‘bystander’ are centered around *how important a person in a photo is* and *the intention of the photographer*. Below, we provide the definitions we used in our study.

Subject: A subject of a photo is a person who is important for the meaning of the photo, e.g., the person was captured intentionally by the photographer.

Bystander: A bystander is a person who is not a subject of the photo and is thus not important for the meaning of the photo, e.g., the person was captured in a photo only because they were in the field of view and was not intentionally captured by the photographer.

The task of the bystander detector (as an ‘observer’ of a photo) is then to infer the importance of a person for the meaning of the photo and the intention of the photographer. But unlike human observers, who can make use of past experience, the detector is constrained to use only the visual data from the photo. Consequently, we turned to identifying a set of visual characteristics or high-level concepts that can be directly extracted or inferred from visual features and are associated with human rationales and decision criteria.

A central concept in the definition of bystander is whether a person is actively participating in an event. Hence, we look for the visual characteristics indicating *intentional posing* for a photo. Other related concepts to this are *being aware* of the photo shooting event and *willingness* to be a part of it. Moreover, we expect someone to look *comfortable* while being photographed if they are intentionally participating. Other visual characteristics signal the *importance of a person for the semantics of the photo* and whether they were *captured deliberately* by the photographer. We hypothesize that humans infer these characteristics from context and the environment, location and size of a person, and interactions among people in the photo. Finally, we are also interested to learn how the photo’s environment (i.e., a public or a private space) affect peoples’ perceptions of subjects and bystanders.

To empirically test the validity of this set of high-level concepts and to identify a set of image features that are associated with these concepts that would be useful as predictors for automatic classification, we conducted a user study. In the study, we asked participants to label people in images as ‘bystanders’ or ‘subjects’ and to provide justification for their labels. Participants also answered questions relating to the high-level concepts described above. In the following subsections, we describe the image set used in the study and the survey questionnaire.

4.2.1 Survey Design

4.2.1.1 Image set

We used images from the *Google open image dataset* [113], which has nearly 9.2 million images of people and other objects taken in unconstrained environments. This image dataset has annotated bounding boxes for objects and object parts along with associated class labels for object categories (such as ‘person’, ‘human head’, and ‘door handle’). Using these class labels, we identified a set of 91,118 images that contain one to five people. Images in the Google dataset were collected from Flickr without using any predefined list of class names or tags [113]. Accordingly, we expect this dataset to reflect natural class statistics about the number of people per photo. Hence, we attempted to keep the distribution of images containing a specific number of people the same as in the original dataset. To use in our study, we randomly sampled 1,307, 615, 318, 206, and 137 images containing one to five people, respectively, totaling to 2,583 images. A ‘stimulus’ in our study is comprised of an image region containing a single person. Hence, an image with one person contributed to one stimulus, an image with two people contributed to two stimuli, and so on, resulting in a total of 5,000 stimuli. If there are N stimuli in an image, we made N copies of it and each copy was pre-processed to draw a rectangular bounding box enclosing one of the N stimuli as shown in Fig. 4.1. This resulted in 5,000 images corresponding to the 5,000 stimuli. From now on, we use the terms ‘image’ and ‘stimulus’ interchangeably.



(a) Image with a single person.
(b) Image with five people
(c) An image where the stimulus is annotated where the stimulus is enclosed by a bounding box.

Figure 4.1: Example stimuli used in our survey.

4.2.1.2 Measurements

In the survey, we asked participants to classify each person in each image as either a ‘subject’ or ‘bystander,’ as well as to provide reasons for their choice. In addition to these, we asked to rate each person according to the ‘high-level concepts’ described above. Details of the survey questions are provided below, where questions 2 to 8 are related to the high-level concepts.

1. Which of the following statements is true for the person inside the green rectangle in the photo? with answer options

i) There is a person with some of the major body parts visible (such as face, head, torso); ii) There is a person but with no major body part visible (e.g., only hands or feet are visible); iii) There is just a depiction/representation of a person but not a real person (e.g., a poster/photo/sculpture of a person); iv) There is something else inside the box; and v) I don’t see any box. This question helps to detect images that were annotated with a ‘person’ label in the original Google image dataset [113] but, in fact, contain some form of depiction of a person, such as a portrait or a sculpture (see Fig. 4.1).

The following questions were asked only if one of the first two options was selected.

2. How would you define the place where the photo was taken? with answer options

i) A public place; ii) A semi-public place; iii) A semi-private place; iv) A private place; and v) Not sure.

- 3. How strongly do you disagree or agree with the following statement: The person inside the green rectangle was aware that s/he was being photographed?** with a 7-point Likert item ranging from *strongly disagree* to *strongly agree*.
- 4. How strongly do you disagree or agree with the following statement: The person inside the green rectangle was actively posing for the photo.** with a 7-point Likert item ranging from *strongly disagree* to *strongly agree*.
- 5. In your opinion, how comfortable was the person with being photographed?** with a 7-point Likert item ranging from *highly uncomfortable* to *highly comfortable*.
- 6. In your opinion, to what extent was the person in the green rectangle unwilling or willing to be in the photo?** with a 5-point Likert item ranging from *completely unwilling* to *completely willing*.
- 7. How strongly do you agree or disagree with the statement: The photographer deliberately intended to capture the person in the green box in this photo?** with a 7-point Likert item ranging from *strongly disagree* to *strongly agree*.
- 8. How strongly do you disagree or agree with the following statement: The person in the green box can be replaced by another random person (similar looking) without changing the purpose of this photo.** with a 7-point Likert item ranging from *strongly disagree* to *strongly agree*. Intuitively, this question asks to rate the ‘importance’ of a person for the semantic meaning of the image. If a person can be replaced without altering the meaning of the image, then s/he has less importance.
- 9. Do you think the person in the green box is a ‘subject’ or a ‘bystander’ in this photo?** with answer options i) Definitely a bystander; ii) Most probably a bystander; iii) Not sure; iv) Most probably a subject; and v) Definitely a subject. This question was accompanied by our definitions of ‘subject’ and ‘bystander’.

10. Depending on the response to the previous question, we asked one of the following three questions: i) **Why do you think the person in the green box is a subject in this photo?** ii) **Why do you think the person in the green box is a bystander in this photo?** iii) **Please describe why do you think it is hard to decide whether the person in the green box is a bystander or a subject in this photo?** Each of these questions could be answered by selecting one or more options that were provided. We curated these options from a previously conducted pilot study where participants answered this question with free-form text responses. The most frequent responses in each case were then provided as options for the main survey along with a text box to provide additional input in case the provided options were not sufficient.

4.2.1.3 Survey implementation

The 5,000 stimuli selected for use in the experiment were ordered and then divided into sets of 50 images, resulting in 100 image sets. This was done such that each set contained a proportionally equal number of stimuli of images containing one to five people. Each survey participant was randomly presented with one of the sets, and each set was presented to at least three participants. The survey was implemented in Qualtrics [6] and advertised on Amazon Mechanical Turk (MTurk) [33]. It was restricted to MTurk workers who spoke English, had been living in the USA for at least five years (to help control for cultural variability [106]), and were at least 18 years old. We further required that workers have a high reputation (above a 95% approval rating on at least 1,000 completed HITs) to ensure data quality [137]. Finally, we used two attention-check questions to filter out inattentive responses [128] (see Appendix A.6).

4.2.1.4 Survey flow

The user study flowed as follows:

1. Consent form with details of the experiment, expected time to finish, and compensation.
2. Instructions on how to respond to the survey questions with a sample image and appropriate responses to the questions.
3. Questions related to the images as described in Section 4.2.1.2 for fifty images.
4. Questions on social media usage and demographics.

4.2.2 Survey Participants and Dataset Labels

4.2.2.1 Demographic characteristics of the participants

Before performing any analysis, we removed data from 45 participants who failed at least one of the attention-check questions. This left us with responses from 387 participants. Of these, 221 (57.4%) identified themselves as male and 164 as female. One hundred and eighty nine (48.8%) participants fell in the age range of 30–49 years, followed by 154 (39.8%) aged 18–29 years. A majority of the participants identified as White (n=242, 62.5%) followed by 82 (21%) as Asian, and 20 (5%) as African American. One hundred and ninety one (49.3%) had earned a Bachelor’s degree, and 71 (18.3%) had some college education. Most of the participants had at least one social media account (n=345, 89.1%), among which only 7% (n=30) indicated that they never share images on those media. Each participant was paid \$7, which was determined through a pilot study where participants were also asked whether they considered the compensation to be fair. Participants were able to pause this survey and resume at a later time, as indicated by the long completion time (> 10 hours) for many of the participants. Therefore we analyzed the response times for the top quartile, which completed the survey in an average of 41 minutes. Thus we estimated that our compensation was in the range of \$10/hour for the work on our survey.⁴

⁴A more conservative estimate yielded about \$8/hour for the top 50%, which took an average of 53 minutes.

4.2.2.2 Final set of images and class labels

For each image, we collected responses from at least three participants. Next, we excluded data for any image for which at least two participants indicated that there was no person in that image (by responding with any one of the last three options for the first question as described in Section 4.2.1.2). This resulted in the removal of 920 images, and the remaining 4,080 images were used in subsequent analyses.⁵ The class label of a person was determined using the mean score for question 9: a positive score was labeled as ‘subject’, a negative score was labeled as ‘bystander’, and zero was labeled as ‘neither’. In this way, we found 2,287 (56.05%) images with the label ‘subject’, 1,515 (37.13%) with ‘bystander’, and 278 (6.8%) with ‘neither’. In this paper, we concentrate on the binary classification task (‘subject’ and ‘bystander’) and exclude the images with the ‘neither’ label. In this final set of images, we have 2,287 (60.15%) ‘subjects’ and 1,515 (39.85%) ‘bystanders’.

4.2.2.3 Feature set

As described in section 4.2.1.2, we asked survey participants to rate each image for several ‘high-level concepts’ (questions 2–8). The responses were converted into numerical values – the ‘neutral’ options (such as ‘neither disagree nor agree’) were assigned a zero score, the left-most options (such as ‘strongly disagree’) were assigned the minimum score (-3 for a 7-point item), and the right-most options (such as ‘strongly agree’) were assigned the maximum score (3 for a 7-point item). Then, for each image, the final value of each concept was determined by computing the mean of the coded scores across the participants. In addition to these, we calculated three other features using the annotation data from the original Google image dataset [113]: size and distance of a person and the total number of people in an image. We estimated the size of a person by calculating the area of the bounding box enclosing the person normalized by total area of the image. The distance refers to the Euclidean distance between the center of the bounding box and the center of the image and can

⁵One of the authors manually checked these images and found that only 9 (0.9%) of them contained people.

be treated as the ‘location’ of a person with respect to the image center. Finally, by counting the number of bounding boxes for each image, we calculated the total number of people in that image. We combined these three features with the set of high-level concepts and refer to this combined set simply as ‘features’ in the subsequent sections.

4.3 Method of Analysis

To understand how humans classify ‘subjects’ and ‘bystanders’ in an image, first, we catalog the most frequently used reasons for the classification (from responses to question 10). Next, we quantify if and how much these reasons are associated with the features as detailed in section 4.2.2.3. Significant association would indicate the relevance of the ‘high-level concepts’ in distinguishing bystander and subject by humans, and serve as a validation for incorporating those concepts in the study. Then, we conducted regression analyses to measure how effective each of the features *individually* are in classifying subject and bystander. Finally, we conducted exploratory factor analysis (EFA) on the whole feature set to surface any underlying constructs that humans use in their reasoning. EFA also helped to group correlated features under a common factor (based on the absolute values of factor loadings), facilitating the selection of a subset of uncorrelated features. Informed by the regression and factor analyses, we identified multiple subsets of features to use as predictors in training classifiers. In the following subsections, we explain each of these steps in more detail.

4.3.1 Quantifying association between human reasoning and features

We employed Spearman’s ρ , which measures the monotonic association between two variables as a correlation measure between the binarized reasons and the real-valued features [61]. Then, for each reason, we grouped the feature values based on whether this reason was used for classification and measured the average of the feature-values in those two groups. We computed Cohen’s d (i.e., the

standardized mean difference or ‘effect-size’) between the two groups and conducted significance tests. A significant difference between the means would signal a feature is indicative of a particular reason.

4.3.2 Measuring predictive-power of individual feature and selecting subset of uncorrelated features

We trained one logistic regression model for each feature (as predictor) to classify ‘subject’ and ‘bystander’. The predictive power of each feature, i.e., how well it alone can predict the class label was assessed by interpreting the model parameters. Our eventual goal is to find a subset of features with (collectively) high predictive power but minimal correlation among them since correlated features can render the model unstable [61]. To find a subset of features that are minimally correlated among themselves but retains maximum variance of the outcome variable, we conducted exploratory factor analysis (EFA) which attempts to discover underlying factors of a set of variables. Below we outline the steps we followed while conducting the factor analysis.

- **Removing collinear variables.** Multiple collinear variables can unduly inflate the variance of each other (i.e. inflate contribution of the variables toward a factor) and so collinear variables should be removed before conducting EFA [220]. First, we standardized the features to remove structural multi-collinearity [1]. Then we tested for multicollinearity using ‘variance inflation factor’ (VIF). We removed features with VIF greater than five [61].
- **Determining the number of factors to extract.** We conducted principal component analysis (PCA) to estimate the amount of variance retained by each component. We decided the number of factors to extract from EFA using a scree plot [61, 157, 220].
- **Extracting and rotating factors.** After removing collinear variables and deciding on the number of factors, we extracted the factors and estimated the factor loading (i.e., correlation between a feature and a factor) of each feature. Finally, we rotated the factors using ‘varimax’

rotation to obtain a simple structure of the factor loadings [157, 220]. The factors become orthogonal (i.e. completely uncorrelated) to each other after the rotation, which makes interpretation easier. Moreover, it helps to group and describe the features, since ideally each feature has a high factor loading for only one factor after the rotation.

Features that are highly correlated among themselves measure the same underlying concept (i.e., factor) and would have high correlation with that factor. Consequently, we grouped the features having high correlation with a single factor into categories describing ‘meaningful’ constructs. This would facilitate in explaining the underlying constructs that are important in the human reasoning process [220]. Additionally, features belonging to one group ideally have low correlation with features belonging to another group. Thus, we identified a subset of minimally correlated features by taking one feature from each group. The collective predictive power of this subset is indicated by how much of the total variance in the full set of variables is retained by the factors.

4.3.3 Developing classifiers using selected feature sets

So far, we have detailed the methods of validating our feature set and identifying subsets of features to be used as predictors. Now, we focus on developing machine learning (ML) models and evaluating their performance. Although we strive to achieve high classification accuracy, we are also interested in learning at what level of abstraction the features have the most predictive power. Thus, we built several classifiers using features at different levels of abstraction, spanning from the raw image to the high-level concepts and evaluated these models by conducting 10-fold cross-validations. Below, we explain these different classifier models.

4.3.3.1 Baseline models

As a baseline model, we started with directly using the cropped images as features to train the classifier. All the cropped images were first resized (256×256 pixels) and then fed into a logistic

regression model. This represents a model trained with the most concrete set of features, i.e., the raw pixel values of the cropped images. Our next classifier is another logistic regression model, trained with higher-level but simple features – the number of people in a photo and the size and the location of each person. This would allow us to investigate if the classification problem can be trivially solved using easily obtainable, simple features.

4.3.3.2 Fine-tuning pre-trained models

Fine-tuning a pre-trained model allows us to transfer learned knowledge in one task to perform some other (often related) task. The process is analogous to how humans use knowledge learned in one context to solve a new problem. Fine-tuning deep learning models has shown great promise in many related problem domains [69, 71, 154, 163]. Here, we fine-tuned ResNet50 [89], which was trained for object detection and recognition on the ImageNet [50] dataset containing more than 14 million images to classify ‘subject’ and ‘bystander’. We chose to use this model since recognizing an object as a ‘person’ is a pre-requisite to classify them as ‘subject’ or ‘bystander’. Hence, the model parameters were pre-trained to optimize recognizing people (and other objects), and we fine-tune it to classify detected people as ‘subject’ or ‘bystander’. To fine-tune this model, we replaced the final layer with a fully connected layer with ‘sigmoid’ activation function. This modified network was re-trained using our (cropped) image dataset. In fine-tuning, we only update the parameters of the last (i.e., newly added) layer, keeping the parameters of all the other layers intact.

4.3.3.3 Models with higher level features

In section 4.3.2, we outlined the process of examining the predictive power of the features and discovering a set of minimally correlated features that best predicts the outcome variable. The feature set includes the high-level concepts, which are not, unfortunately, directly derivable from the image data with currently available machine learning models. We attempt to overcome this barrier by utilizing existing ML models to extract features that we believe to be good proxies for

the high level concepts. We then train two classifiers by – 1) training directly with these proxy features and 2) following a *two-step* classification pipeline by first training regression models with the proxy features to predict the high-level concepts and then using the *predicted* values of the high-level concepts to train the final classifier. Below, we detail what proxy features we extracted and how.

- **Human related features.** The ResNet50 [89] model was trained to categorize objects (including people) in images. We feed the cropped images of people in our dataset in the pre-trained model and extract the output of the second-to-last layer of the network to be used as features for our classifier. Since the original RestNet50 network uses these features in the last layer to assign an object to the appropriate class, and the class in our case is ‘person’, the features are presumably useful in distinguishing people from other objects. In other words, these features are useful in detecting people, which is a prerequisite for classifying a person as a subject or bystander.
- **Body-pose related features.** We used OpenPose [36] to estimate body-pose of a person, which attempts to detect 18 regions (or joints) of a human body (such as nose, ears, and knees), and outputs detected joints along with detection confidence. We used the confidence scores, which indicate how clearly different body parts of a person are visible in an image, as feature values. Additionally, for each pair of neighboring joints (e.g., right shoulder and right elbow), we computed the angle between a line connecting these joints and the horizontal axis. Collectively, these angles suggest the pose and the orientation of the body. These features were extracted from OpenPose [36] using the cropped images of each person. But in our dataset, some cropped images contain body parts of more than one person (see Fig. 4.2), and OpenPose attempts to detect all of them. Since in our case a single stimulus (i.e. cropped image) is associated with one person, we needed to single out the pose features for that person only. For example, Fig. 4.2a shows a cropped image where two people are visible, but the

original image was cropped according to the bounding box for the person at the right side of the cropped image. Although OpenPose detects body parts for both people, we need this information only for the person with whom this image is associated (in this case the person at the right side), since the pose features will be used to classify that person only. We use a simple heuristic to solve this problem – a cropped image is associated with the most centrally-located person. With this heuristic, when a body part (such as nose) was detected more than once, we retain information about the part that is closest to the center of the cropped image. Fig. 4.2b shows the result of body part detection using this mechanism.

- **Emotion features estimated from facial expression.** We extracted scores for seven emotions: ‘angry’, ‘disgusted’, ‘fearful’, ‘happy’, ‘sad’, ‘surprised’, and ‘neutral’. Intuitively, these features might be good proxies for ‘awareness’, ‘comfort’, and ‘willingness’ of a person. To obtain emotion features, we first extracted faces from the cropped images using a face detection model [96]. If two people appear in each other’s cropped images, each of them will be positioned in a more central location of the cropped image associated with them and will be detected with higher accuracy and confidence by the face detection algorithm. Hence, in cases where a cropped image contains multiple people, we retained the face that was detected with the highest confidence. After detection, the faces were extracted and fed into a facial expression recognition model [123]. Using facial features, this model estimates the probabilities of each of the seven emotions. We used these probability values as features.

4.3.4 Comparing ML models with humans

One way to investigate how well the ML models perform compared to humans is to compare how much human annotators agree among themselves with the model accuracy. Computing agreement statistics, however, require all annotators to label the same set of images, which is infeasible in this case. Hence, instead of agreement among the annotators, we computed what percentage of



(a) The colored dots show the body joints of the two people originally detected.

(b) Result of removing duplicate body joints based on the distance from image center.

Figure 4.2: Detecting and refining body joints.

annotators agreed with the final class label of an image. Recall that the final class label was decided by taking the mean of the scores for ‘subject’ and ‘bystander’ (provided by the survey participants). For example, if two participants labeled someone as ‘most probably a subject’ (coded value = 1), and a third participant labeled that person as ‘most probably a bystander’ (coded value = -1), then the mean score is 0.3. Hence, the final label of that person would be ‘subject’, where 67% annotators agreed with this label. We grouped the images based on what percentage of the annotators agreed with its label. We then used these groups individually to train classifiers and test their performance for image sets with varying degrees of agreement.

4.3.5 Test Dataset

We assessed the performance and robustness of the models created with the above-mentioned steps with 10-fold cross-validation using non-overlapping train-test splits of the Google dataset [113]. To evaluate how well our approach generalizes to different datasets, we conducted additional analysis (using the model trained on the Google dataset) on an independent dataset consisting of 600 images

sampled from the *Common Objects in COntext* (COCO) dataset [126]. COCO contains a total of 2.5 million labeled instances in 328,000 images of complex everyday scenes containing common objects in their natural context and has been used in numerous studies as a benchmark for object recognition and scene understanding. We randomly sampled roughly equal number of photos with one to five people totalling to 600 samples of individual person. Using this sample, survey data was collected and analyzed in the same way as explained above, but participants from the previous study were not allowed to take this survey. After pre-processing the survey data, we found that 354 (59%) and 246 (41%) people in the images were labeled as ‘subject’ and ‘bystander’, respectively.

4.4 Findings

4.4.1 How Humans Classify ‘Subjects’ and ‘Bystanders’?

The most frequently used reasons for labeling a person as a ‘subject’ or a ‘bystander’ by the survey participants are shown in Tables 4.1 and 4.2. For ‘subjects’, the top four reasons involve visual characteristics of the individual person under consideration (Table 4.1). Intuitively, these reasons are related with the visual features we extracted from the images and collected using survey responses (we quantify these associations and present the results in the next section). For example, ‘being in focus’ with size and location of a person, ‘taking a large space’ with size, and ‘being the only person’ and ‘activity of the person being the subject matter of the image’ with importance of the person for the semantic of the image or if the person can be replaced without altering the semantic content. The last three reasons consider overall image context and visual similarities of the person in question with other people in the same image (Table 4.1).

Similarly, the most frequently selected reason for labeling a person as a ‘bystander’ (Table 4.2) is ‘not focusing on the person’, which is associated with the size and location of that person in the image. The second most frequent reason is ‘caught by chance’, which again relates to if that person is important for the image or can be replaced. Reasons 4 and 5 were chosen when participants thought

Table 4.1: Most frequent reasons found in the pilot study for classifying a person as a *Subject* and how many times each of them was selected in the main study.

#	Reason	Frequency
1	This photo is focused on this person.	5091
2	This photo is about what this person was doing.	4700
3	This is the only person in the photo.	2740
4	This person is taking a large space in the photo.	2425
5	This person was doing the same activity as other subject(s) in this photo.	2357
6	This person was interacting with other subject(s) in this photo.	1715
7	The appearance of this person is similar to other subject(s) of this photo.	1644

no person was a subject of the image or there was no specific subject at all. The other reasons consider overall image content and visual similarity and interactions of the person in question with other people in the image (Table 4.2). These results indicate that the human decision process for this classification task considers visual characteristics of the person in question (e.g. size) as well as other people in the image (e.g. interaction among people in the image). This process also involves understanding the overall semantic meaning of the image (e.g., someone was captured by chance and not relevant for the image) and background knowledge (e.g., if two people have similar visual features or are performing the same activity, then they should belong to the same class). Such rich inferential knowledge is not available in images. Since our ultimate goal is to build classifiers that only use the images as input, we investigate the relationships of the human rationale with visual features that can be extracted from the image.

4.4.2 Association Between Human-reasoning and the Features

4.4.2.1 How Well are the ‘High-level Concepts’ and the ‘Features’ Associated with the Reasons Humans Used?

The correlations between the features and the reasons for specific labels and the standardized differences between the means in feature values when a specific rationale was used or not used for

Table 4.2: Most frequent reasons found in the pilot study for classifying a person as a *Bystander* and how many times each of them was selected in the main study.

#	Reason	Frequency
1	This photo is not focused on this person.	3553
2	This person just happened to be there when the photo was taken.	2480
3	The activity of this person is similar to other bystander(s) in this photo.	1758
4	Object(s) other than people are the subject(s) of this photo.	1644
5	Appearance of this person is similar to other bystanders in this photo.	1278
6	There is no specific subject in this photo.	849
7	This person is interacting with other bystander(s).	755
8	This person is blocked by other people/object.	567
9	Appearance of this person is different than other subjects in this photo.	537
10	The activity of this person is different than other subjects(s) in this photo.	466

labeling are presented in Tables 4.3 and 4.4.⁶ Significant correlation coefficients and differences in group means suggest an association between the features and the rationales. As an example, the positive correlation coefficient of 0.19 indicates that when participants thought that *the photo was focused on a person*, they also tended to agree more on the assertion that that person was *posing* for the photo. Similarly, the (standardized) difference between the means of the ‘Posing’ feature when participants used the reason *the photo was focused on that person* to label a person as a subject versus when they did not use that reason is 0.42.⁷ This implies that being ‘in-focus’ of a photo is related to the concept of ‘posing’ for that photo. Associations among the other reasons and high-level concepts can be similarly interpreted.

Table 4.3: Correlation coefficients and effect sizes between the visual features and the reasons for classifying a person as a *subject*. All coefficients and effect-sizes are significant at $p < .001$ level.

Feature	Spearman ρ	Cohen’s d
<i>This photo is focused on this person</i>		
Awareness	0.17	0.36

⁶Since the features are related to individual people and do not capture the interactions among people or the overall contexts of the images, we present results only for the reasons referring to individual persons.

⁷Cohen’s $d=0.2$, 0.5 , and 0.8 are considered to be a ‘small’, ‘medium’, and ‘large’ effect size respectively [42].

Pose	0.19	0.42
Comfort	0.15	0.30
Willingness	0.15	0.30
Replaceable	-0.20	-0.39
Size	0.35	0.69
Distance	-0.29	-0.63
Number of people	-0.37	-0.82

This person is taking a large space in the photo.

Awareness	0.11	0.22
Comfort	0.11	0.24
Willingness	0.12	0.25
Replaceable	-0.20	-0.43
Size	0.38	0.83
Distance	-0.19	-0.43
Number of people	-0.20	-0.44

This is the only person in this photo.

Awareness	0.11	0.21
Pose	0.10	0.21
Replaceable	-0.12	-0.24
Size	0.27	0.65
Distance	-0.23	-0.47
Number of people	-0.61	-1.33

Table 4.4: Correlation coefficients and effect sizes between the visual features and the reasons for classifying a person as a *bystander*. All coefficients and effect-sizes are significant at $p < .001$ level.

Feature	Spearman ρ	Cohen d
<i>This photo is not focused on this person.</i>		
Awareness	-0.25	-0.59
Pose	-0.31	-0.77
Comfort	-0.25	-0.49
Willingness	-0.26	-0.52
Replaceable	0.16	0.31
Photo place	-0.22	-0.52
Size	-0.20	-0.44
Distance	0.21	0.46
<i>This person just happened to be there when the photo was taken</i>		
Awareness	-0.34	-0.70
Pose	-0.36	-0.72
Comfort	-0.19	-0.33
Willingness	-0.22	-0.41
Replaceable	0.27	0.50
Photo place	-0.24	-0.49
Size	-0.23	-0.37
Distance	0.13	0.26
<i>This person is blocked by other people or object.</i>		
Awareness	-0.15	-0.46

Pose	-0.17	-0.54
Comfort	-0.11	-0.29
Willingness	-0.12	-0.37
Replaceable	0.14	0.38

4.4.2.2 Identifying subsets of *uncorrelated* features that are effective in distinguishing ‘subject’ and ‘bystander’

First, we trained separate classifier models for each feature as a predictor to asses how well each of them can individually distinguish between a ‘subject’ and a ‘bystander’. We report the detailed results in Appendix A.1. In summary, all of the features (described in Section 4.2.2.3) were found to be significantly associated with the outcome (i.e., subject and bystander), but the magnitude of the predictive power varied across features. We also found that almost all pairs of features have medium to high correlations between them (Appendix A.2). Hence, we conducted EFA to discover uncorrelated feature sets.

As outlined in Section 4.3, first we calculated VIF to detect multicollinearity (Table A.3). Among the features, ‘Awareness’ has the highest VIF of 5.8 (and a corresponding $R^2 > .8$ in the regression model), indicating that this feature can be predicted almost perfectly using a linear combination of other features. This is also apparent in the pairwise correlations among the features (see Appendix A.2), where ‘Awareness’ is highly correlated with most of the other features, making it redundant. Removal of this feature resulted in a drop of VIF for every other feature below 5, suggesting a reduction in multicollinearity in the system (re-calculated VIF are shown in the second column of Table A.3).

With the remaining features, we conducted PCA to find out the appropriate number of factors to extract [220]. The point of inflexion [220] in the Scree plot (Fig. 4.3) after the second factor suggests the extraction of two factors, which jointly retain approximately 60% of the total variance

in the data. Fig. 4.4 exhibits the factor loadings of each feature after a ‘varimax’ rotation [61]. We omitted the features with factor loadings less than 0.32 [157].⁸ A feature is associated with the factor with which it has a higher loading than the other, and the features associated with the same factor were grouped together to form descriptive categories [220]. More specifically, ‘Pose’, ‘Comfort’, and ‘Willingness’ were grouped together under the category ‘visual appearance’ of a person. This grouping makes sense intuitively as well since all three variables refer to the body orientation and facial expression of a person. Similarly, ‘Size’, ‘Distance’, and ‘Number of people’ collectively represent ‘how prominent’ the person is in the photo.⁹ Finally, ‘Replaceable’ has almost equal loadings on the two factors and, hence, was not assigned to any group. Intuitively, it suggests how ‘important’ a person is for the semantic meaning of the image, which depends on both the ‘visual appearance’ and ‘prominence’ of a person.

Upon grouping the features that are highly correlated among themselves, we now select a subset of features by picking one feature from each group (‘Pose’ and ‘Size’, respectively) and the two features (‘Replaceable’, and ‘Photographer’s intention’) that do not belong to any group.¹⁰ ‘Replaceable’, and ‘Photographer’s intention’. Results from a linear regression model trained with this feature set is shown in Table 4.5. This model has a better fit with the data ($R^2 = 0.53$) than any of the models trained with individual features (Table A.1). But this model utilizes ground truth data about ‘Pose’, ‘Replaceable’, and ‘Photographer’s intention’ obtained from the user study, which can not be extracted directly from the image data. In the next section, we present classification results using different feature sets produced from the images.

⁸The location of a person did not have high enough correlation with any of the factors. Hence, it was not used in subsequent analysis.

⁹Although ‘Size’ appears to be far from the others, this is because it has positive association with ‘Factor2’, while the rest have negative association. This is also intuitive, since as the ‘Number of people’ and ‘Distance’ increase, size should decrease.

¹⁰We experimented with different combinations of features from these two groups and obtained comparable results.

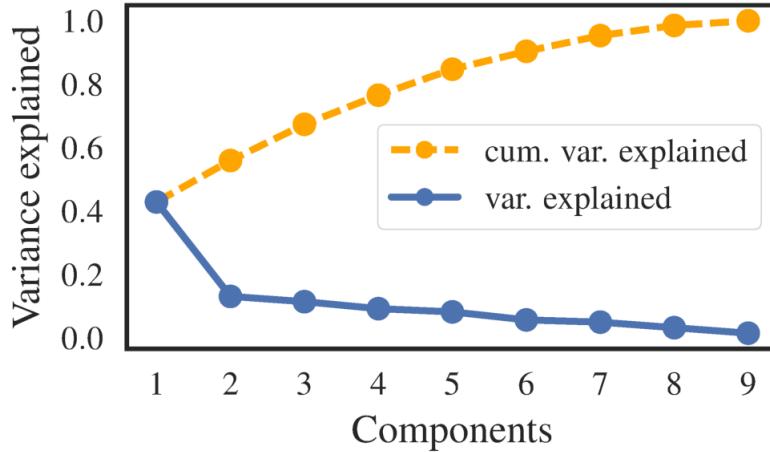


Figure 4.3: Scree plot showing *proportions of variance* and *cumulative proportion of variance* explained by each component extracted using PCA.

Table 4.5: Effectiveness of the selected features to classify ‘subject’ and ‘bystander’. The columns show odds-ratios and their 95% confidence intervals for each feature. All $p < 0.0001$.

	Odds Ratio	[95% CI]
Pose	2.50	[2.17, 2.88]
Replaceable	0.13	[0.11, 0.15]
Size	1.91	[1.64, 2.22]
Photographer’s intention	0.56	[0.49, 0.63]

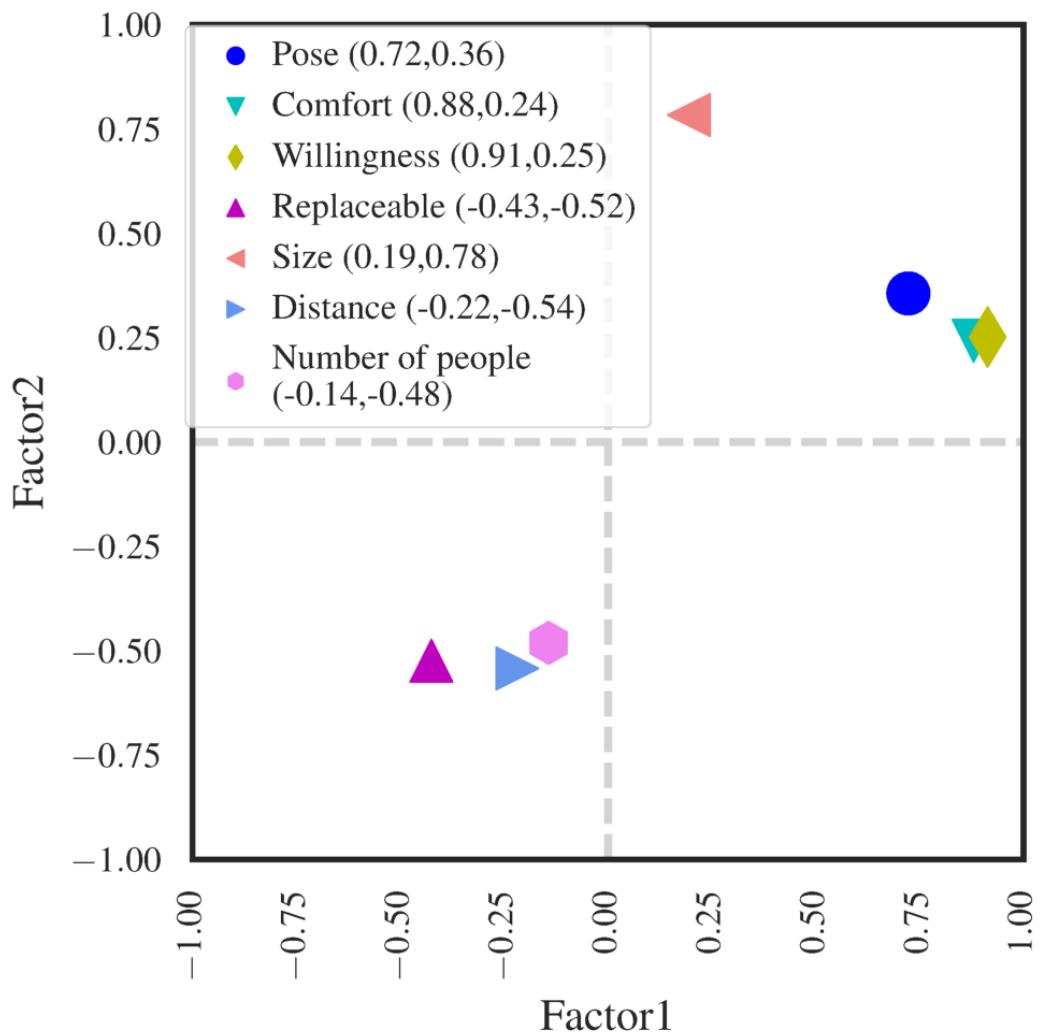


Figure 4.4: Factor loadings of the features across the two extracted factors. The numeric values of the loadings are displayed within braces with the legend.

4.4.3 Machine Learning Models to Predict ‘Subject’ and ‘Bystander’

Table 4.6 shows means and standard-deviations for classification accuracy using different feature sets (including the model using ground truth high-level concepts). Fig. 4.5 shows the corresponding Receiver Operating Characteristic (ROC) plots for each case generated from 10-fold cross-validation. Using the cropped images as features has the lowest mean accuracy of 66%. Using the simple features – ‘Size’, ‘Distance’, and ‘Number of people’ – yielded mean accuracy of 76%, a 15% increase than using raw image data. We see a corresponding increase in the area under the curve (AUC) measure in Fig. 4.5. Fine-tuning the pre-trained ResNet [89] model did not improve the accuracy any further (Table 4.6).

Using ground truth values of the high-level concepts, combined with the ‘Size’ feature increased the accuracy by more than 12% (mean accuracy $86\% \pm 0.04$ and AUC 93%). Next, we employ the proxy features of these high-level concepts as detailed in Section 4.3.3.3 and obtained a mean classification accuracy of 78%, a small increase from the model using simple features. Finally, we use the *predicted* values of the high-level concepts using the *proxy features* and obtained a mean accuracy of 85% and corresponding AUC of 93%, *which is similar to the results obtained using ground truth values* of the high-level concepts (details on prediction accuracy are provided in Appendix A.3). We obtained similar results using different subsets of predicted features, indicating that predictors in the same set contain repeated information and do not add any new predictive power, which again validates our EFA analysis.

From these results, we see that features at a higher level of abstraction yield better classification accuracy. The raw image, despite having all the information present in any feature derived from it, performs noticeably worse than even the simple feature set. Similarly, *predicted* values of the high-level concepts performed better than the proxy features they were predicted from. Although the proxy features presumably contain more information than any feature predicted from them, the high-level concepts are more likely to contain information relevant for distinguishing subjects

Table 4.6: Mean and standard deviation of accuracy for classification using different feature sets across 10-fold cross validation.

Features	Accuracy	
	Mean	SD
<i>Cropped image</i>	66%	0.03
<i>Size, distance, and number of people</i>	76%	0.01
<i>Fine-tuning ResNet</i>	77%	0.02
<i>ResNet, Pose, and Facial expression features</i>	78%	0.03
<i>Size and ground truth Pose, Replaceable, Photographer’s intention</i>	86%	0.04
<i>Size and predicted Pose, Replaceable, Photographer’s intention</i>	85%	0.02

and bystanders in a more concise manner and with less noise.

4.4.4 Comparing ML Models with Humans

The percentages of agreement among the annotators and the number of images for each percentage are presented in Appendix A.4. All annotators agreed on the final label for only 1,309 (34%) images, and for 1,308 (34%) images there were agreements among two-third of the annotators. For these two groups of images, we train and evaluate classifiers following the two-step procedure.¹¹ For a 10-fold cross validation, the mean classification accuracy were 80%(± 0.03) and 93%(± 0.02), respectively for these two groups (The corresponding ROC plots are shown in Appendix A.5). Considering the fact that these two models were trained using much smaller sets of images than before, they achieved remarkably high accuracy even for the images with only 67% agreement among human annotators.

4.4.5 Accuracy on the COCO Dataset

For the 600 images sampled from COCO [126], our model (trained on the Google data set) achieved an overall classification accuracy of 84.3%. To compare the accuracy with humans, we again divided these images based on how many of the annotators agreed with the final label. We found that 354

¹¹We did not perform similar analyses for images with lower than 67% agreement because of insufficient training data. We had only 400 such images.

(59%) images had 100% agreement, while 168 (28%) images had 67% agreement. For these two subsets, our model achieved 91.2% and 78.6% classification accuracy, respectively. The results of this extended analysis are consistent with the results with the Google dataset and provide strong evidence for the generalization of our approach and trained models.

4.5 Limitations and Discussion

Photography as art. We must note that just because bystanders *can* be detected does not mean that they *should* be removed or redacted from images, or that a particular bystander should necessarily exert control over the image. There are legitimate reasons for bystanders to be retained in images, ranging from photo-journalism to art. The questions of image ownership and the right to privacy of bystanders are complicated and depend on contextual, cultural, and legal factors. Nevertheless, in many circumstances, owners of photos may voluntarily be willing to redact images out of a sense of ‘propriety’ and concern about bystanders [95]. For example, Anthony et al. discuss how people routinely engage in behaviors to respect the privacy of others [18]. Other work seeks to make privacy ‘fun’ by encouraging owners of photos to apply stickers or redactions on bystanders [84, 85]. Our work on detecting bystanders should thus be seen as a necessary building block of larger automated frameworks that consider further action on photos.

People detection. For the Google dataset [113], we used manually annotated bounding boxes to locate people and extracted features from these cropped images. Results may differ if people were instead detected automatically, but we do not expect large deviations since computer vision can detect and segment people with close to human-level performance [172].

Annotators. All of our survey participants were U.S. residents (although the images used had no such restriction); future work could consider cross-cultural studies. We used three annotators per image under the assumption that unanimous agreement among three independent observers is a strong signal that a given person is indeed a ‘bystander’ or ‘subject’. We expect that requiring

agreement among more annotators would slightly reduce the size of the dataset but also increase the accuracy of our algorithm for that dataset, as any ambiguity is further reduced. Overall, three annotators struck a reasonable balance for such labeling.

Dataset. We considered images containing one to five people for practical reasons. In our labeled data, we noticed that as the number of people per image grows, fewer of them are labeled as subjects. This indicates that, as one might expect, images with large numbers of people typically contain crowds in public places, with no particular subject. Including such images would result in an imbalanced dataset and ultimately a biased model. We hypothesize that classifying subjects and bystanders in such images would be easier than in images with fewer people since people usually have smaller size and are not centrally located (size and location features have significant positive and negative correlations with being a subject) in those images. Finally, we observed that beyond some threshold, people with smaller size are much harder to recognize. Thus, we expect that our algorithm will not only scale to images with larger crowds but will yield better classification accuracy.

Feature relationships. Another limitation of our work is that we use features only from individual people as predictors. However, as our user study uncovered, relationships and interactions among people in an image also play important roles in the categorization of *subject* vs. *bystander*. For example, some participants labeled a person as a ‘bystander’ because they “looked similar to” or “were doing the same activity as” another bystander. Future work should investigate classifiers that incorporate these inter-personal relationships.

Use of additional metadata. Our goal in this paper is to propose a general-purpose bystander detector using visual features alone, to make it as widely applicable as possible, including on social media platforms, image-hosting cloud servers, and photo-taking devices. We expect that accuracy can be increased using contextual information available in any specific domain, e.g., using image captions, one’s friend list in a social network, and location of the photo. In the future, we plan to

explore the use of domain-specific information.

4.6 Conclusion

Photographs often inadvertently contain bystanders whose privacy can be put at risk by harming their social and professional personas. Existing technical solutions to detect and remove bystanders rely on people broadcasting their privacy preferences as well as identifying information – an undue burden on the victims of privacy violations. We attempt to tackle the challenging problem of detecting bystanders automatically so that they can be removed or obfuscated without proactive action. Our user study to understand the nuanced concepts of what makes a ‘subject’ vs. ‘bystander’ in a photo unveiled intuitive *high-level concepts* that humans use to distinguish between the two. With extensive experimentation, we discovered visual features that can be used to infer those concepts and assessed their predictive power. Finally, we trained machine learning models using selected subsets of those concepts as features and evaluated their performance. Our best classifier yields high accuracy even for the images in which the roles of subjects and bystanders are not very clear to human annotators. Since our system is fully automated, and solely based on image data, it does not require any additional setup and can be used for any past, present, and future images, we believe that it has the potential to protect bystanders’ privacy at scale.

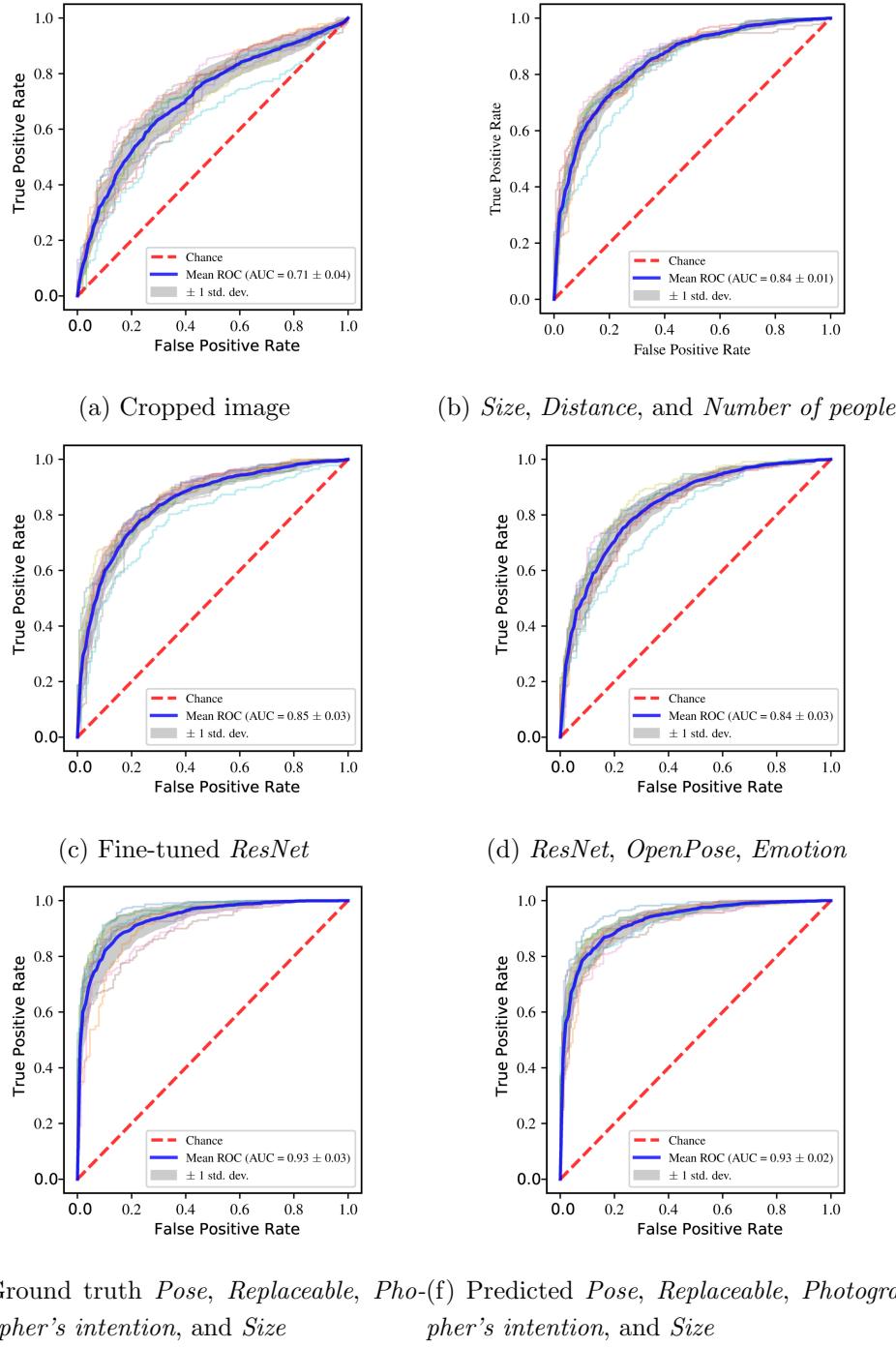


Figure 4.5: Receiver operating characteristic (ROC) plots for classifier models using different feature sets.

CHAPTER 5

Evaluating Image Filters in Terms of Privacy-Protection Capability and Usability

In chapter 4, we introduced an automated system to detect bystanders in images. Other researchers have proposed machine learning based models to detect several other types of privacy-sensitive objects in images, such as specific places [201] and electronic screens [110]. After detecting sensitive photo-contents, one approach to reduce privacy risks is to obfuscate scene elements and image filters (e.g., blurring, pixelating, and silhouette) have been used for this purpose for a long time. This chapter describes a systematic study evaluating such filters in terms of their ability to properly obscure the intended information and preserve utility in transformed images for the viewers.

This study was done in collaboration with Eman Hassan, Yifang Li, Roberto Hoyle, David Crandall, and Apu Kapadia. Findings from this study were published as “Viewer Experience of Obscuring Scene Elements in Photos to Enhance Privacy” in CHI’18 [84].

5.1 Introduction

Among the billions of photos that are shared online each day [108], many of them are taken in public places and often contain bystanders. When shared online, these photos violate bystanders’ privacy by revealing their identity and location to an unbounded number of people. Beyond identity and location researchers have identified many other attributes of people and other objects that are considered to be privacy sensitive (e.g., an embarrassing facial expression) [12,14,41,77,95,118,189]. Indeed, even photo-sharers have expressed their own privacy concerns over these attributes [12,14, 41,118,189]. We seek to understand the efficacy of image filters in obscuring such sensitive contents at the *attribute level*.

Filtering regions of an image presents a trade-off between privacy and utility; these transformations need to be aggressive enough to remove or obscure private information, but not so aggressive

that they destroy the value of sharing the image. For example, transformations such as blurring and pixelation can be used to redact portions of an image [24, 31, 77, 111, 125, 146]. However, much of this work does not consider the potential negative impact on image aesthetics or utility, and much of it focuses on obfuscating faces or bodies, not on various other scene elements that may also raise privacy concerns (e.g., monitors and financial documents). While some work considers how these transformations affect the user experience [76, 125] or studies particular transformations of objects [86], we believe a systematic study is needed on how well various transformations balance concealing private content with preserving image value for a human viewer.

In this work we examine how obfuscating ‘objects’ affects various ‘attributes’ of those objects that a viewer can perceive. We present the findings of an experimental study conducted on Amazon Mechanical Turk¹ ($N=570$) on the effects of five different transforms (*masking, blurring, pixelation, edge detection, and silhouetting*) on both privacy and user experience (including visual aesthetics and satisfaction) for scenarios that previous studies have identified as important for privacy [12, 14, 41, 77, 95, 118, 189].

We find that it is possible to protect selected regions within an image while preserving utility and aesthetics. We also find that different filters work better at protecting different attributes within images, and provide quantitative information to guide future applications.

5.2 Experiment

We conducted an experiment to study the effectiveness of several image obfuscation methods (see Table 5.1)² designed to conceal objects and different properties of them, as well as how well these obfuscation methods retain image utility. We included twenty different scenarios in which we varied objects and their properties, as described in Table 5.2. Each of these scenarios had one of twelve different conditions, each using a different method and/or degree of obfuscation. Participants were

¹<https://www.mturk.com/>

²We did not include actual photos that were used in the survey in Table 5.1 due to copyright issues. To obtain the photos please contact one of the authors.

randomly assigned to one of the twelve filter conditions (between subjects). Each participant was then presented with all 20 scenarios in random order (within subjects).

5.2.1 Measurements

In the experiment we asked five questions for each scenario:

What is the object (or property of the object) depicted in the image? This question varied slightly based on the scenario; in Table 5.2 we summarize the specific questions we used. Participants were asked to select from multiple-choice options consisting of the most common answers given in the pilot study, which had a free-form text box. Answers were marked either correct or incorrect. A green bounding box was overlaid surrounding the objects of interest to ease locating them only for this question, and later removed for subsequent questions (as described below) for the same scenario.

How confident do you feel that you correctly answered the previous question? This question used a 7-point Likert scale.

For the next three questions, we asked the participants whether they agreed or disagreed with the following statements.

The photo provides sufficient information. This item (also on a 7-point Likert scale) is adapted from the ‘information quality scale’ [179], which measures “the satisfaction of users who directly interact with the computer for a specific application” [46]. We adapted “Does the photo provide sufficient information,” which loads onto the “content” factor and was strongly correlated with questions “is the system successful?” and “are you satisfied with the system?”

The photo is satisfying. We adapted this item from the validated ‘image appeal scale’ [46], which is the extent to which images are perceived as “appropriate and aligned to user expectations, satisfying, or interesting... and goes beyond aesthetics or the attractiveness.” Specifically, this selected item measures the participants’ overall ‘satisfaction’ with the image after the alterations,

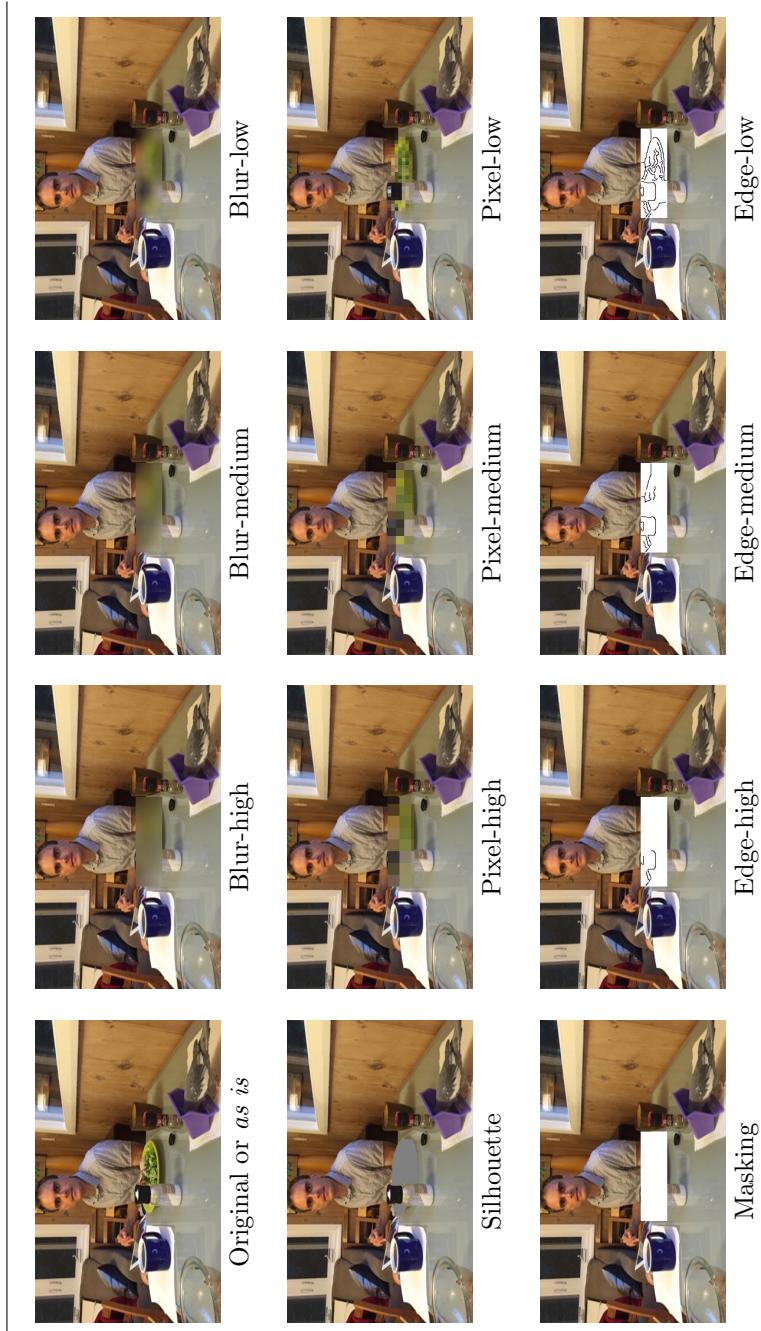


Table 5.1: Results of applying different filters to obscure food.

Scenario	Question
Activity	What is the person inside the green rectangle doing?
Age	What is the age of the person inside the green rectangle?
Document class	What is the object inside the green rectangle?
Document text	What is the text inside the green rectangle?
Document type	What type of document (e.g. notebook, paper) is inside the green rectangle?
Dress	What type of clothing is the person inside the green rectangle wearing?
Ethnicity	What is the ethnicity of the person inside the green rectangle?
Expression	What is the facial expression of the person inside the green rectangle?
Food	What type of food is inside the green rectangle?
Gender	What is the gender of the person inside the green rectangle?
Hair	How long is the hair of the person inside the green rectangle?
Indoor	Was the following photo taken indoor or outside?
Indoor specific	What type of indoor place (e.g., library, concert hall) is shown in the following photo?
Laundry	What is the object inside the green rectangle?
Messy room	How well organized or messy is the place shown in the photo?
Monitor app.	What application is displayed on the computer monitor inside the green rectangle?
Monitor class	What is the object inside the green rectangle?
Monitor text	What is the text inside the green rectangle?
Outdoor	Was the following photo taken indoors or outside?
Outdoor specific	What type of outdoor place (e.g., field, street) is shown in the following photo?

Table 5.2: Scenarios and the recognition questions used in the survey.

as also measured by Li et al. [125] when obscuring faces and bodies. A 7-point Likert was used.

This photo looks visually appealing. To frame this item, we asked participants to “*Imagine a friend of yours shares this photo on a social networking site, such as Facebook,*” and was also measured on a 7-point Likert scale.

5.2.2 Scene selection

Our scenarios are representative of the objects and properties about which privacy concerns were expressed in prior studies [12, 14, 41, 77, 95, 118, 189]. Through these scenarios we capture peoples’ concerns related to privacy of information (e.g., leaking text from financial documents or computer screens), impression management based on appearance (e.g., facial expression, hair style), activities (e.g., using social media during work hours), and living conditions (e.g., messy room, eating habits).

5.2.3 Obfuscation Methods

Along with the *as is* (unaltered) condition as a control, we used five primary types of obfuscations: *Blurring*, *Pixelating*, *Edge* (i.e., line drawing), *Masking*, and *Silhouette*. This selection was informed by prior studies according to the appropriateness for the research questions we seek to answer. Earlier studies on *blurring* and *pixelating* were limited primarily to facial identity protection [125], and found that these filters are well accepted by users but not effective when applied at a level that preserves photo utility [76, 148]. We thus wanted to determine their effectiveness to conceal other objects and properties. *Masking* has been found to be effective to protect identity but hides masked photo content completely [125]; we study its effect when applied on objects that are small or not the main subject matter of the photo. *Silhouette* is interesting because it preserves shape, which we hypothesized may be useful to retain an object’s identity but remove finer details that might contain private information. On the other hand, *edge* preserves shape and some internal details and may be useful in cases where finer control is required.

While the *masking* and *silhouette* filters are binary, either completely obscuring the original object or not, the other three have continuous-valued filter parameters. Applying *blur* and *pixelating* filters with low parameter values generates output images that are similar to the originals, while increasing values cause the filtered image regions to be more aggressively obscured. The *edge* filter parameter controls a threshold on edge strength, with higher values removing all but the strongest lines while lower values retain more detail.

These leveled filters might be effective in obscuring different types of information at different parameter values. For example, blurring with an aggressive filter value may be able to completely obscure an object such as a computer monitor, whereas blurring with a milder value might only obscure details (e.g., text on a monitor screen) but not the object itself. To study these effects, we included the *masking*, *silhouette*, and *blur* filters with ‘high,’ ‘medium,’ and ‘low’ levels of aggressiveness in our experiment. These values were determined through a smaller user study

(more details are provided in the supplementary materials) in which we developed a tool and showed different images at different levels to participants. For each respondent, the filter levels were decreased until he or she was able to determine the identity of the object, and in turn (for the next level) detect lower-level features in the image. The levels across all participants and images were averaged. The ‘high’ level was chosen as the average level for high-level details plus one standard deviation, and likewise ‘low’ was equal to the average for low-level details plus one standard deviation. The ‘medium’ level was chosen as the average for high-level details. These three filters with three levels each, along with the *masking*, *silhouette*, and *as-is* (no filtering), make up the twelve obfuscation methods in our study, and are summarized in Table 5.1.

5.2.4 Collecting Images

For each scenario we used different image sets, so that any image for one scenario would not reveal answers about any other scenario. Each set consisted of ten images collected from the internet. Using more than one image for each scenario allows us to incorporate some controlled variability and draw more useful conclusions from the study than for a single image. At the same time, we were careful to select images that had consistent image properties, such as brightness and object size, in each scenario. In particular, we tried to follow the following guidelines as closely as possible:

1. The quality, illumination, and size should be as consistent as possible across all images.
2. For any particular scenario, all five images should have a similar number of people and/or other objects with similar distribution and orientation.
3. For any particular scenario, the object of interest (e.g. face) should be of comparable size across all five images.
4. The object of interest should not be the focus of the image; e.g., when monitors are the object of interest, the monitor should not be in the center or ‘too large’ compared to other objects

in the image. We are interested in cases where information is leaked through objects that are not the main subject matter and may go unnoticed.

5. It should not be possible to easily identify the object or property of interest from scene context, such as other objects or properties (e.g., computer monitors next to adjacent keyboards, type of indoor place like library from collection of bookshelves, food from the logo of the restaurant or other food in the vicinity, and so on).

From these images, we further sub-sampled the five images for each scenario which were most consistent with the guidelines. Finally, we scaled the images to be consistently sized. Our pilot study did not reveal any systematic large differences in identification accuracy for any specific image in a scenario.

5.2.5 Organization of the Survey

The survey instrument was organized as follows:

1. Consent form.
2. Questions about which (if any) social media services the participant uses, how frequently they share images using those services, and four demographic questions.
3. Instructions on how to answer the survey questions along with a sample image either in *as is* condition, or a filter that was randomly selected and applied on a predefined region.
4. Twenty scenarios, presented in a random order (within subjects), each with five questions in a specified order. Each scenario presented one of five random images modified by one of the twelve obfuscation methods (between subjects — each participant was assigned to a single transform condition, e.g., *Blur-medium*, was selected at random and fixed for the participant).

The experiment was implemented in Qualtrics and is included in Appendix B.1.

5.2.6 Ethical Considerations

The study was approved by the Institutional Review Board at Indiana University.

5.2.7 Recruitment, Compensation, and Validation

The study was advertised on Amazon Mechanical Turk as an “Image Transformation Study.” Participants were required to live and have resided in the United States for at least five years, in order to reduce cultural variability [106]. To ensure higher data quality [137], we restricted to MTurk workers with high reputation (above 95% approval rating on at least 1000 completed HITs). They were also required to be at least 18 years of age; studying photo obfuscation preferences and experiences of teenagers could be an interesting direction for future work. The average time to complete the survey was around 20 minutes, and respondents were compensated US\$2.50 upon completion of the study. We paid all 725 respondents who completed the study, but eliminated participants from our sample if they failed any of the three attention-check questions, leaving 570 participants in our final sample.

5.2.8 Pilot Study

We first performed a pilot study with N=45 respondents, also administered via Amazon Mechanical Turk, but respondents were compensated \$3.00. Data from this pilot study was used to estimate the sample size required to produce statistically significant findings through a power analysis. Moreover, the top five free-form text responses for the recognition questions were used as the multiple-choice options (instead of a text field) in the final study. We acknowledge the concern that providing a fixed number of choices can make picking the correct option easier than answering correctly in free-form text. In the pilot, however, we found that participants were already using contextual information present in the photos and for any particular question the number of different replies were less than ten. Furthermore, our experimental setup provides insights through the relative

changes observed across conditions.

For each scenario, we used the most common response in the *as is* condition during the pilot study as the *correct answer* in the final study. The pilot also helped us to test for unforeseen variability within our images which might lead users to misidentify the objects of interest, but did not find any.

5.3 Findings

5.3.1 Demographic Information

Among the 570 participants, 324 (56.8%) described themselves as male, and 172 (30.1%) were non-white. 197 (34.6%) were aged between 18–29 years, 303 (53.2%) between 30–49 years, 63 (11.1%) between 50–64 years, and 7 (1.2%) above 65 years. The highest level of education attained was high school for 203 (35.6%) participants, undergraduate degree for 306 (53.7%), Masters for 49 (8.68%), and Ph.D. or professional degree for 12 (2.08%). All participants reported using at least one social media service, and 293 (51.4%) reported sharing images using social media at least “a few times” a week.

5.3.2 Recognition Accuracy

In order to characterize how well filters obscure potentially private information we look at two metrics. First, we measured “accuracy” as the participants’ ability to recognize objects and properties in transformed images by simply computing the fraction of correct responses. We analyzed the responses using Fisher’s exact test, where we compared the accuracy of each filter with the accuracy of the *as is* condition, and present the results (recognition rate, p-value, and effect size) in Table 5.3. We applied the Bonferroni correction for these tests (i.e., for each row (filter) of the table, we corrected for 11 hypothesis tests against the baseline filter). Next, we also looked at the effect size to measuring the effectiveness of the filter over the as-is baseline. For Fisher’s exact test,

the effect size is the ratio of the odds of being correct in a treatment condition (i.e., filter) to the odds of being correct in the control condition (i.e., *as is*), so lower effect sizes correspond to lower odds of being correct when a filter is applied. In our case, this helps us determine how effectively the filter prevents recognition. We designate filters as *effective* when the recognition accuracy is less than 50% and the effect size is less than 0.05,³ and *somewhat effective* when accuracy is less than 50% and effect size less than 0.1. For example, with 50% and 95.2% recognition accuracies for filtered and unfiltered conditions respectively, the odds ratio is 0.05, indicating that *effective* filters drastically lower the odds of recognition success.

In general, we observed that *blurring*, *pixelating*, and *edge* filters at low and medium levels are effective in protecting specific or minor details, such as text, but fail to obscure general properties such as whether an object is a document or monitor. These filters are almost always ineffective even in the strongest levels for scenarios that require obscuring the entire image. In contrast, *masking* is effective at obscuring objects (as well as almost all other scenarios) and *silhouette* is mostly effective for objects and attributes that cannot be recognized from shape (e.g., *ethnicity*). Below we describe the findings for each filter in more detail.

5.3.2.1 Blurring

We found that *blurring* at a low level is only effective in obscuring *activity*, *gender*, *document type*, *document text*, and *monitor text* (Table 5.3). In addition, mid-level blurring can prevent recognition of *monitor application* and *specific indoor environment*. On the other hand, a high level of *blurring* is effective in all scenarios except *expression*, *monitor class*, *general outdoor environment*, and *messy room*, and for *ethnicity*, *specific outdoor environment*, and *food*, it is only somewhat effective. In summary, *blurring* is not effective at protecting properties related to *Environment*, food, and

³A threshold for recognition accuracy in the filtered condition is required to get a meaningful effect size. Otherwise when accuracy is 100% for the unfiltered condition, the odds ratio will be zero for any accuracy in the filtered condition and the filter will appear as *effective* even if it fails to prevent recognition (i.e., high recognition accuracy for the filtered object). Also we select these values to ensure that for effective filters, the recognition probability is less than 50% chance and without any filter the recognition probability is close to certainty (100%)

laundry (specially at low and medium levels), but effective for other scenarios at high and medium levels.

5.3.2.2 Pixelating

High and medium levels of *pixelating* perform similarly to corresponding levels of *blurring* across all scenarios except for human attributes (e.g., *facial expression*, *dress*) where *pixelating* seems better than *blurring* (Table 5.3). On the other hand, a low level of *pixelating* is only effective for *activity*, *document text*, and *monitor text*, and performs worse than a low level of *blurring* in other cases. But a low level of *pixelating* preserves more information and generates more visually appealing photos compared to *blurring* (and other filters) as discussed in the section on 5.3.4, and so might be more desirable than other filters when effective.

5.3.2.3 Edge

Similar to *blurring* and *pixelating*, the *edge* filter becomes more effective as the filter parameter becomes more aggressive. However, unlike the other two, a high level *edge* filter is at least somewhat effective for all the scenarios related to *document* and *computer monitor* (Table 5.3). *Edge* is also effective at obscuring *food* at both high and medium levels, and effective for *laundry* even at a low level. In short, *edge* seems to be more effective than *blur* and *pixelate* when the object to be obfuscated has irregular shape and/or internal texture that produces *noise-like curves* when the filter is applied.

5.3.2.4 Silhouette and Masking

Silhouette and *masking* filters are similar in the sense that they completely remove or replace the filtered region. But since *silhouette* preserves shape information, we expected it to be effective for objects and properties that cannot be recognized by boundaries. We found *silhouette* effective for all scenarios except *hair*, *monitor class*, and *food* (Table 5.3), which we expected, but also for *age*,

which was surprising because high levels of *blurring* and *pixelating* are effective in this case. We believe this is because a person’s body shape and posture can be strong cues for *age*, and *silhouette* reveals information about these. Another interesting finding is that *masking* fails to protect *facial expression* despite being effective in all other cases. This is because our definition of effectiveness required an effect size (ratio of recognition accuracies in filtered to unfiltered conditions) less than 0.05, but the accuracy on the *as-is* condition was already so low that *masking* did not create enough *additional* confusion to be considered effective. Note that we did not test environmental and text related scenarios with *silhouette* because it is not clear how this transform would be applied in these cases different from masking.

5.3.3 Recognition Confidence

In general, the mean confidence value for the *as is* condition was the highest for all scenarios as expected.⁴ Next we analyzed participants’ confidence levels separately for correct and incorrect answers. For incorrect identification, there were no significant differences in confidence levels for any filter across any scenario. When identified correctly, generally the mean confidence levels were higher than when identified incorrectly. Moreover, we found that confidence levels vary significantly for different filters for most of the scenarios, but interestingly, for difficult and/or confusing scenarios (such as *document text*, *expression*, and *hair*), we did not find any significant difference in confidence for any filter. This indicates the inherent ambiguity involved in identifying these properties of images and participants were not very confident about their (correct) identification.

5.3.4 Photo Utility

We next analyzed how well filters preserved the perceived *utility* of images, using the three questions from the survey on whether an image “provides sufficient information,” “is satisfying,” and “looks

⁴For *specific indoor environment*, *monitor text*, and *specific outdoor environment*, the highest values were for *masking*, *pixelate-medium*, and *pixelate-low* respectively, although the differences with the *as is* condition were not statistically significant.

visually appealing.”

5.3.4.1 Information Sufficiency

Viewers’ perceptions of information sufficiency in an image is affected by obfuscation [125]. By performing an overall Kruskal-Wallis test for all conditions in each scenario, we found significant variations, but the actual H-statistic values differ for different scenarios, and for any particular scenario not all obfuscation methods have a significant effect on information sufficiency, as shown in Table 5.4. The least effective filters in terms of recognition accuracy also have the least damaging effect on information sufficiency. Conversely, filtered images that obscure sensitive information also tend to lack sufficient information even at low levels (such as *edge* for activity and text). Unsurprisingly, the highest level of these filters and *masking* significantly destroy information content in most of the scenarios. We examine the relationship between information sufficiency and recognition accuracy in more depth in the next section.

To allow us to draw more general conclusions, we categorized the scenarios into five groups: Human (activity, age, dress, expression, ethnicity, gender, hair), Monitor (monitor class, monitor application, monitor text), Document (document class, document type, document text), Environment (indoor, indoor specific, outdoor, outdoor specific, messy room), and Other (laundry, food). Figure 5.1 presents mean responses for the information sufficiency question (in terms of the 7-point Likert scale) for each filter and scenario group. We noticed that for scenarios where only small portions of images are obfuscated, all filters have comparable mean values (Figure 5.1). For Human properties, *pixel-low* has the highest mean value among all filters, followed by *blur-low*. For Document, all levels of *blur* and *pixel* (except *pixel-low*) along with *masking* have lower values than the average value of the scale (3.5), while *silhouette* and *edge-low* have values close to *as is*. This is probably due to the fact that documents have rigid shapes which are better preserved by *silhouette* and *edge* filters compared to others. For monitor attributes, *pixel-low* and *silhouette* retain more

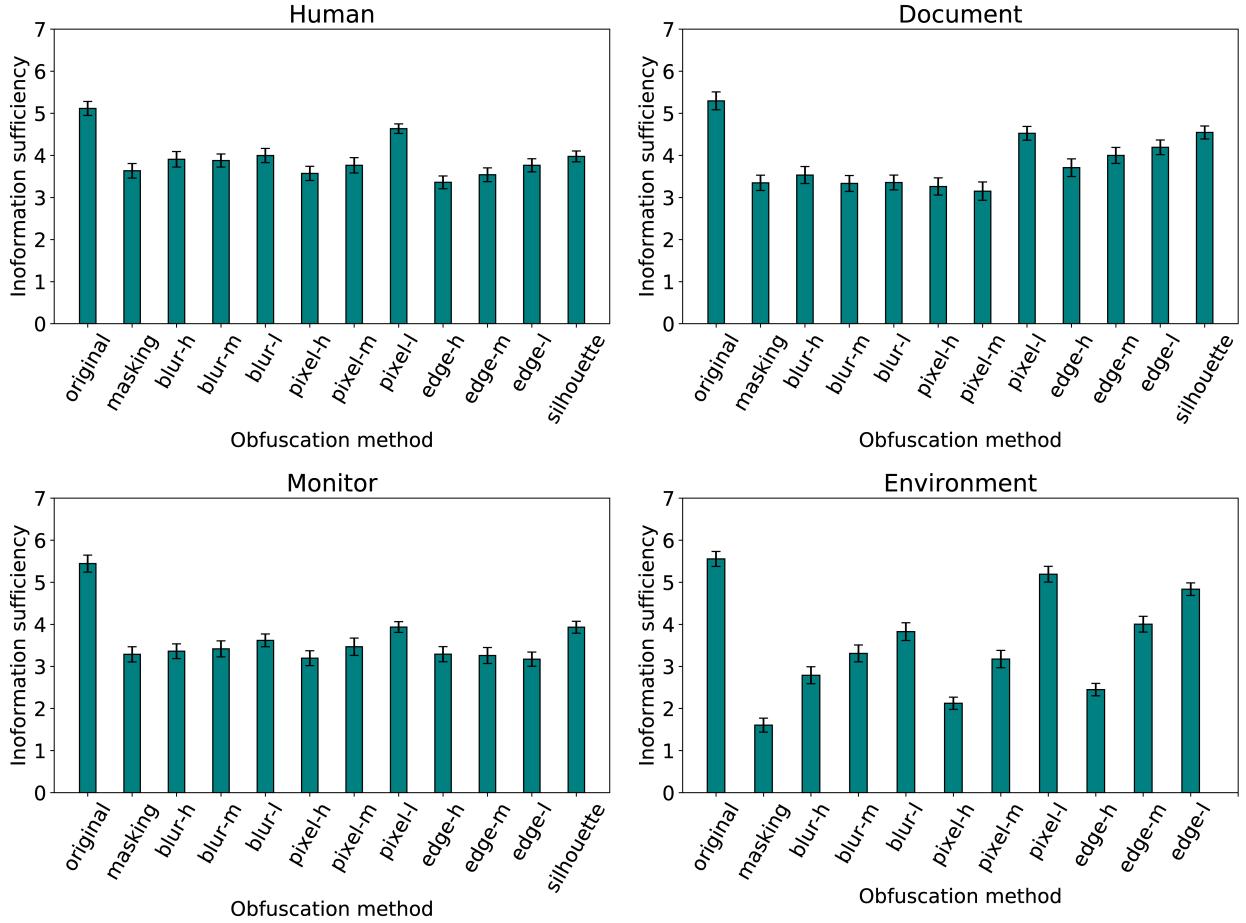


Figure 5.1: Information sufficiency across scenario groups and filters, in terms of mean values and standard error.

information compared to others, while for environment scenarios where we obfuscate the whole image, we observe large differences in mean values of *pixel-low* and *edge-low* compared with others. In summary, the weakest filters (e.g. *pixelate-low*) preserve the most information, and information content is inversely proportional to the filter strength, conforming to prior studies [125], and is proportional to the area of the filtered region.

5.3.4.2 Photo Satisfaction and Visual Aesthetics

We observe that less aggressive (and thus often less effective) filters such as *blur low* and *pixelate low* generate images that are more satisfactory and visually appealing. However, satisfaction and

aesthetics also depend on the size of the obfuscated region. For full image obfuscation (such as indoor/outdoor environment) and full human body obfuscation (such as dress and ethnicity), both satisfaction and aesthetics are hampered. Interestingly, while satisfaction and visual appeal are highly correlated (0.66 correlation), information sufficiency is much less correlated with both of these (correlations 0.44 and 0.25), suggesting that reduced information is not necessarily always accompanied by lower satisfaction (as we discuss in the next section). We also observe similar mean values across filters for these two measures both for individual scenarios and grouped scenarios, so we only include the plot of photo satisfaction of grouped scenarios, in Figure 5.2. Similarly, we study the relationship of recognition accuracy with only visual aesthetics in detail in the next section.

5.3.5 Privacy-Utility Trade-off

Figure 5.3 visualizes the trade-off between obscuring sensitive information and retaining image utility, using scatter plots of information sufficiency (y-axis) versus recognition accuracy (x-axis) for grouped scenarios. We see a roughly linear, positive correlation between detection accuracy and information sufficiency. This suggests that viewers of an image perceive it to be lacking information when they fail to recognize objects or properties of interest in the image. For groups *Human*, *Document*, and *Monitor*, we see clusters of filters in the left region of the plots. We find that *blur-high* for *Human*, and *silhouette* for all categories except *Environment* might strike the best balance between privacy and perceived information sufficiency. For *Environment*, where the whole image is obfuscated, the points form a diagonal line, indicating a clear trade-off between privacy protection and information content of images. In this case, a medium level of *blur* and *pixelate* provides a reasonable balance between recognition accuracy and the amount of information retained in obfuscated photos.

Photo satisfaction and visual aesthetics were closely correlated (0.66 correlation), so we only

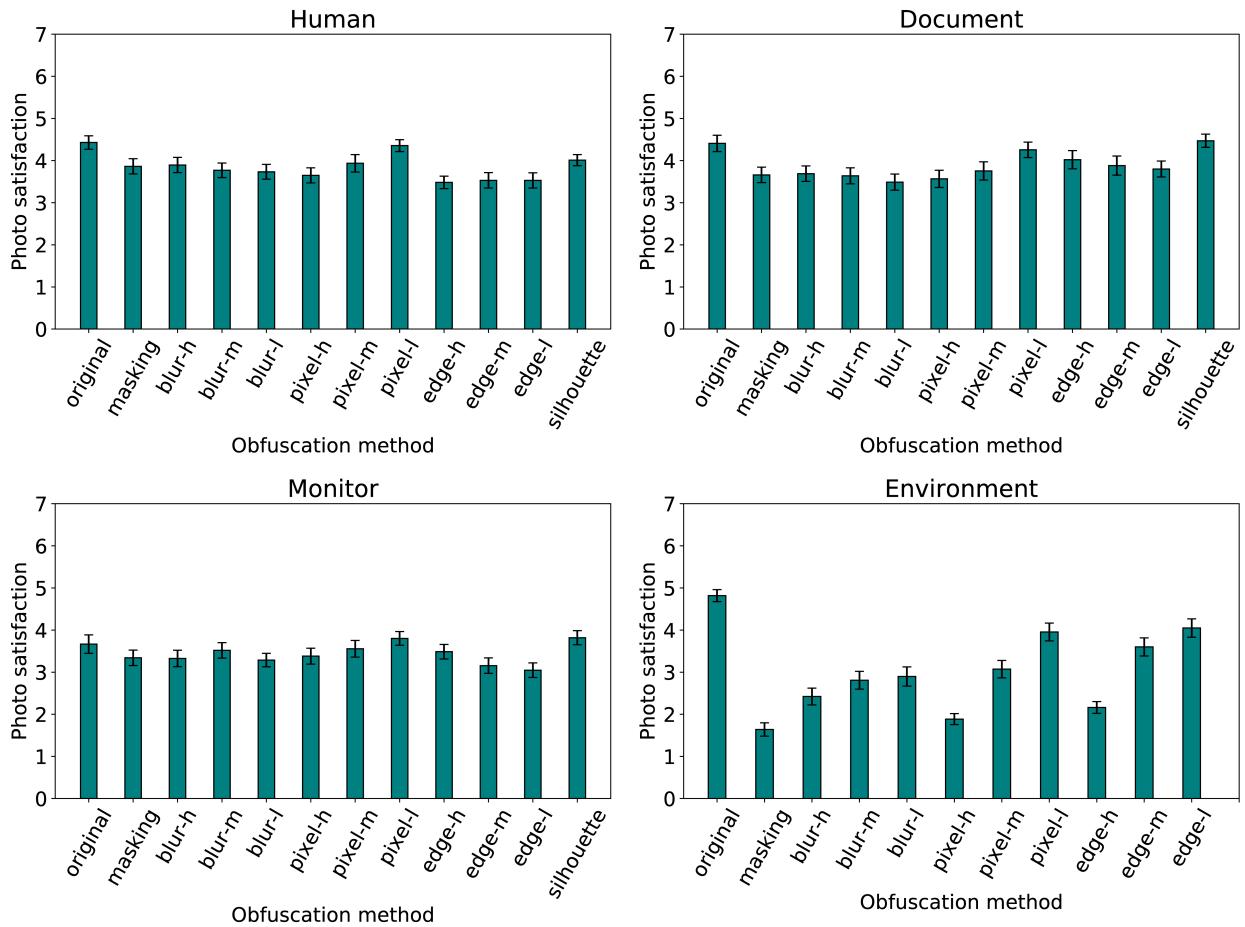


Figure 5.2: Photo satisfaction across scenario groups and filters, in terms of mean values and standard error.

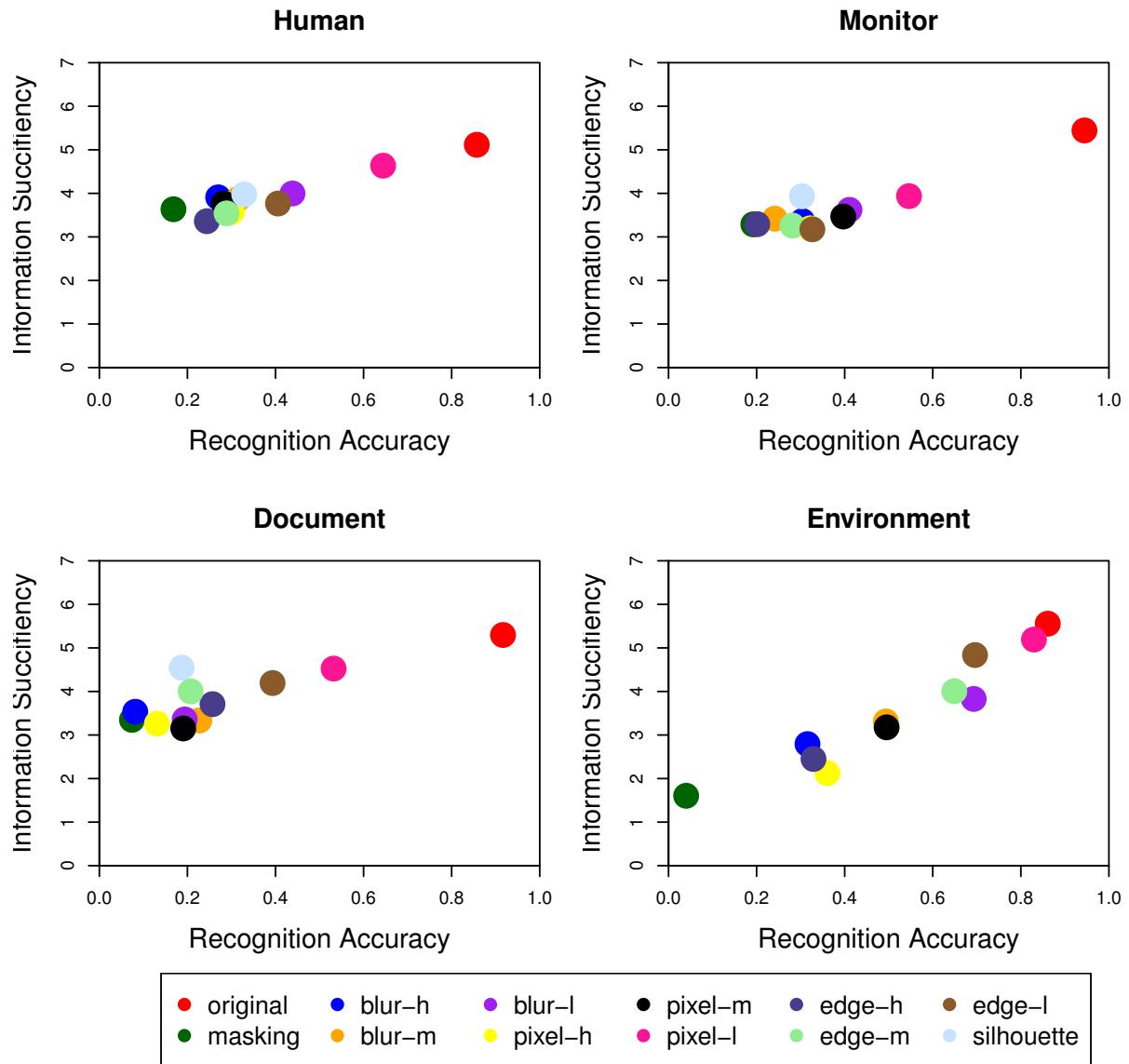


Figure 5.3: Trade-off between protecting against information leaks and information sufficiency across filters, in terms of recognition accuracy (x-axis) and mean information sufficiency (y-axis). Note that *Silhouette* was not studied for any property related to *Environment*.

discuss visual aesthetics. Figure 5.4 compares recognition accuracy and visual aesthetics. We see that for *Environment* the filters are distributed diagonally, meaning that there is a clear trade-off between privacy and visual aesthetics. But for other scene categories there are filters that both protect privacy and leave the filtered image visually appealing: such as *silhouette* for all categories; *pixelate-high* and *blur-medium* for *Monitor*; *blur-high*, *pixelate-high*, and surprisingly, *masking* for *Human*.

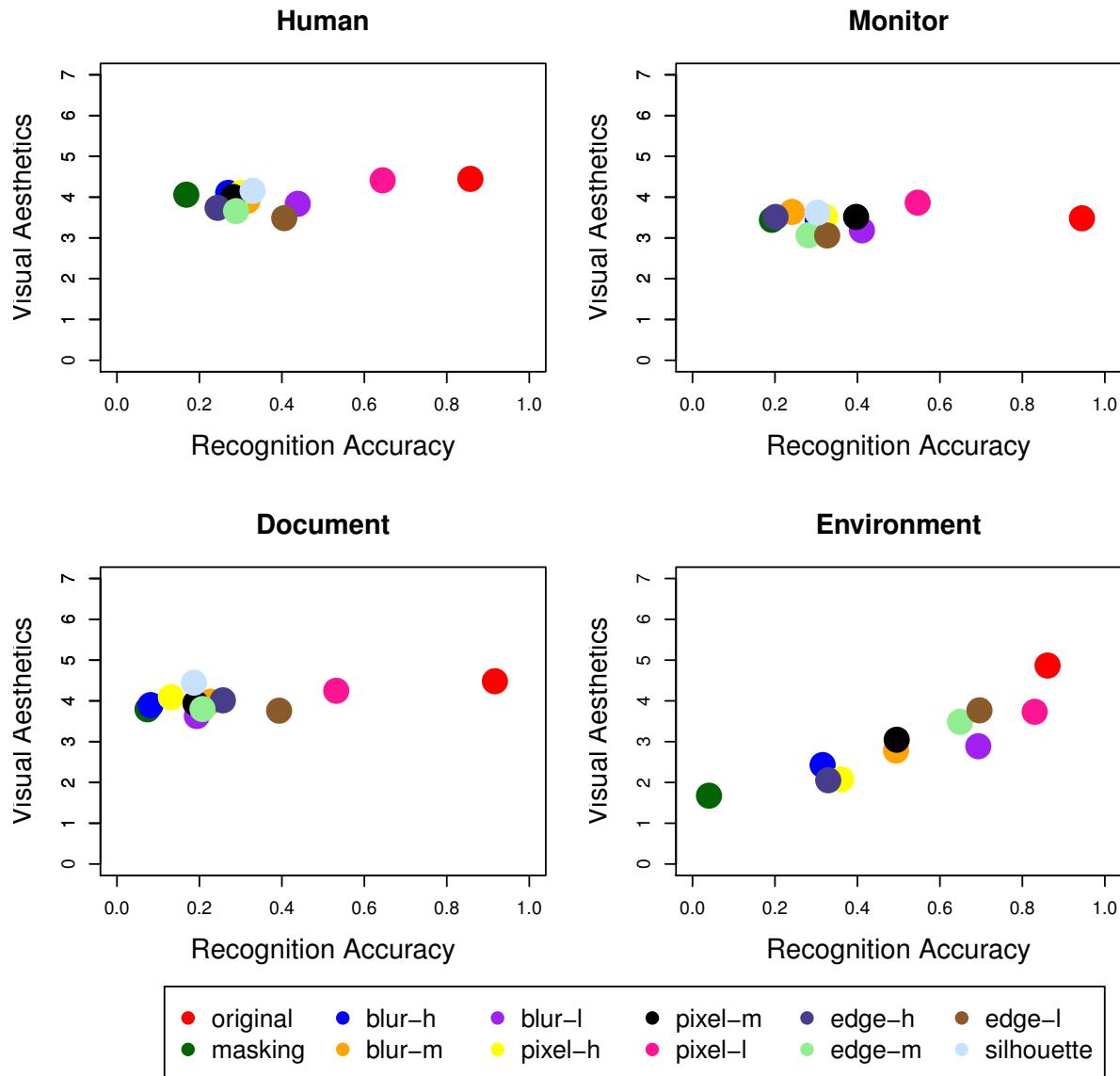


Figure 5.4: Trade-off between protecting against information leaks and aesthetics across filters, in terms of recognition accuracy (x-axis) and mean visual aesthetics (y-axis). Note that *Silhouette* was not studied for any property related to *Environment*.

5.4 Discussion

We now examine the implications, future work possibilities, and limitations of our study.

5.4.1 Privacy vs. Utility

In general, our findings are in line with earlier work [31, 125]: stronger filters increase perceived privacy and decrease perceived information content, satisfaction, and aesthetics. This is especially true for scenarios with specific answers (e.g., dress and gender) or when the whole photo is filtered. However, when a filtered object is small and/or not an integral part of the scene but nevertheless potentially privacy sensitive (e.g., a document), perceived information content and visual aesthetics remain high. This indicates that enhancing the privacy of images does not always result in a reduced user experience. For example, at one extreme, silhouetting objects provides complete privacy for all object attributes other than the type of object, but results in high scores for visual aesthetics. We also found that weaker filters and levels (such as *blur-low*) have little effect on obscuring people, monitors, and documents across a range of situations, again confirming prior findings. This demonstrates that all filters are not equivalent, and different solutions may be more appropriate for different user needs and content types.

5.4.2 Effectiveness of Filters Throughout Categories

The effectiveness of obscuring information for the leveled filters is highly correlated with the specificity of the information that the filter is intended to obscure. At their most aggressive levels, these filters can prevent leaking major details (such as the photo environment or gender of a person), but at medium and low levels are effective only in protecting minor details and specific information (e.g., text or age). On the other hand, since *silhouette* preserves the shape of object boundaries but redacts everything else, we expect it to fail to protect information leakage only when the information can be inferred from the shape of the boundary (such as *food* and *monitor class*). For

objects with rigid boundaries, *silhouette* is as effective as *masking*, which is the most effective filter we found.

For subjective and difficult scenarios (as indicated by low recognition accuracy in the *as-is* condition in Table 5.3) such as *facial expression*, *age*, *messy room*, and *hair length*, all filters seem to be less effective than scenarios with straightforward answers (such as *text*). But note that effect size is a relative measure with respect to the *as-is* condition, so that low baseline accuracy worsens the effect size, meaning applying any filter does not add much confusion.

5.4.3 Edge Detection Side Effects

The *edge* filter behaves differently than the other leveled filters: while *blur* and *pixelate* produce an image very similar to the original at their lowest levels, *edge* always produces a binary “edge map” (as shown in Table 5.1). At its most aggressive, *edge* shows only the most prominent lines, and as the parameter is decreased, progressively weaker lines are revealed. Intuitively, the edge map contains more information for lower values of the parameter, but in some cases, detection accuracy actually *decreased* for lower parameter values (e.g., gender, hair). We speculate that in some scenarios, such as *document type* and *monitor type*, applying the edge filter at a high level leaves the obfuscated region with lines that amplify prominent rectangular objects that are distinctive of these objects. Meanwhile, the abundance of distracting edges at lower filter values makes it more difficult to correctly identify objects. In effect, *edge* applied with a low parameter actually increases noise, and can make it harder to infer information when the filtered regions have too much detail.

5.4.4 Implications and Practical Applications

We expect that our work will shed light on how to transform elements within an image to preserve privacy. Our work shows, as one might expect, that there is no ‘one size fits all’ filter for obscuring scene elements. Depending on the application, different objects can be obscured with custom

filters, and our work makes the first step at trying to characterize how different filters applied at varying levels affect what is concealed and revealed about objects. These findings may improve user acceptability and privacy protection applications such as transforming real-time video streams [86] by selecting the transformation type in an object-dependent way. Our work also offers insight for mobile applications such as VizWiz [22], which allow people with visual impairments to take photos of their environment and ask questions about it to crowd workers, social media friends, or automated applications. While these applications have tremendous potential to help people with visual impairments, there are also severe privacy risks, since users do not necessarily know what their photos contain. Our findings provide a way to transform images so that only the image elements required to answer a particular question are retained. Finally, photos shared via social media can be privacy sensitive for their owner and/or bystanders, and our findings can be integrated into privacy preserving image sharing frameworks such as PuPPIeS [88], and combined with proposed methods to automatically detect sensitive contents in photos [207].

5.4.5 Human vs. Computer Viewers

As discussed in the Introduction, this work does not consider computer-vision based attacks. While certain types of transforms can be defeated by computer vision better than humans, other transforms (such as those applied to CAPTCHAs) defeat computer vision algorithms but not humans. Our work considers human viewers of images and our findings can be interpreted in conjunction with research on computer vision based attacks, based on the application and adversary model, in particular considering whether or not information needs to be revealed to human viewers and the impact of the transforms on human experience. Future work can further study the trade-offs of computer-vision based adversaries.

5.4.6 Limitations

We speculate that filters covering only background or foreground elements or of different sizes may exhibit different results. We made attempts to control for this by making sure that the main, centered, foreground object was not the one that was filtered, and that the filtered area was not so big so as to occlude most of the image or so small that it was hard to spot. However, we have not systematically studied how the size or location of obscured regions within a photo affects how they are perceived; this is a worthwhile direction for future work.

Another limitation is that we only compare the performance of each filter against an *as is* baseline, as opposed to other myriad possible comparisons. Due to the number of conditions in our study, we struck a balance between the resources needed and the number of insights that could be drawn with sufficient statistical significance.

Finally, this study was conducted on Amazon Mechanical Turk, whose user demographics are not representative of the general U.S. population and are known to be more privacy conscious [105, 174]. Nevertheless, we attempt to measure the information loss through objective measures. Other studies have shown that such crowd platforms are a reasonable choice for studying user experience [128].

5.5 Conclusions

Our work sheds light on the effects of applying various types of image transforms to scene elements in an image. In particular we studied the relative trade-offs between privacy (revealing and concealing selective attributes of objects) and utility (the visual aesthetics and user satisfaction of the image) of five different image transforms and show that while in some cases a clear privacy vs. utility trade-off is realized, in other scenarios a high degree of privacy can be attained while retaining utility. Our work also contributes significantly to the existing literature by examining these trade-offs for a range of objects and their attributes, whereas previous work had focused largely on obscuring people and faces. We hope our work spurs further research on studying the relative trade-offs of

image transformations for enhanced privacy without (significantly) degrading the user experience of the viewers.

	<i>Original</i>	<i>Masking</i>	<i>Blur</i>			<i>Pixelate</i>			<i>Edge</i>			<i>Silhouette</i>
	high	medium	low	high	medium	low	high	medium	high	medium	low	
Human												
Activity	97%	8% <i>H</i>	15% <i>H</i>	6% <i>H</i>	16% <i>H</i>	14% <i>H</i>	11% <i>H</i>	48% <i>H</i>	14% <i>H</i>	35% <i>H</i>	64% <i>N</i>	10% <i>H</i>
Age	88%	17% <i>H</i>	26% <i>H</i>	36% <i>M</i>	62% <i>N</i>	42% <i>M</i>	42% <i>M</i>	85% <i>N</i>	37% <i>M</i>	35% <i>M</i>	50% <i>N</i>	50% <i>N</i>
Dress	100%	28% <i>H</i>	44% <i>H</i>	51% <i>N</i>	81% <i>N</i>	44% <i>H</i>	38% <i>H</i>	93% <i>N</i>	27% <i>H</i>	44% <i>H</i>	44% <i>H</i>	48% <i>H</i>
Ethnicity	88%	17% <i>H</i>	28% <i>H</i>	59% <i>N</i>	65% <i>N</i>	48% <i>N</i>	52% <i>N</i>	78% <i>N</i>	8% <i>H</i>	24% <i>H</i>	46% <i>N</i>	25% <i>H</i>
Expression	72%	22% <i>N</i>	28% <i>N</i>	21% <i>N</i>	13% <i>M</i>	19% <i>M</i>	11% <i>M</i>	38% <i>N</i>	31% <i>N</i>	13% <i>M</i>	16% <i>M</i>	18% <i>M</i>
Gender	100%	20% <i>H</i>	24% <i>H</i>	25% <i>M</i>	46% <i>H</i>	28% <i>H</i>	33% <i>H</i>	72% <i>N</i>	37% <i>H</i>	26% <i>H</i>	52% <i>N</i>	43% <i>H</i>
Hair	52%	2% <i>H</i>	20% <i>H</i>	21% <i>N</i>	20% <i>N</i>	12% <i>N</i>	7% <i>M</i>	34% <i>N</i>	14% <i>N</i>	22% <i>N</i>	12% <i>N</i>	33% <i>N</i>
Document												
Document class	86%	13% <i>H</i>	17% <i>H</i>	42% <i>N</i>	27% <i>M</i>	26% <i>M</i>	40% <i>N</i>	65% <i>N</i>	27% <i>M</i>	24% <i>M</i>	48% <i>M</i>	10% <i>H</i>
Document type	97%	8% <i>H</i>	6% <i>H</i>	23% <i>H</i>	25% <i>H</i>	12% <i>H</i>	14% <i>H</i>	72% <i>N</i>	33% <i>H</i>	26% <i>H</i>	54% <i>M</i>	45% <i>H</i>
Document text	91%	0% <i>H</i>	0% <i>H</i>	2% <i>H</i>	4% <i>H</i>	0% <i>H</i>	2% <i>H</i>	21% <i>H</i>	16% <i>H</i>	11% <i>H</i>	16% <i>H</i>	—
Monitor												
Monitor class	100%	42% <i>H</i>	68% <i>N</i>	57% <i>N</i>	88% <i>N</i>	60% <i>N</i>	71% <i>N</i>	89% <i>N</i>	45% <i>H</i>	55% <i>N</i>	64% <i>N</i>	81% <i>N</i>
Monitor app.	88%	15% <i>H</i>	22% <i>H</i>	14% <i>H</i>	34% <i>M</i>	35% <i>M</i>	45% <i>N</i>	74% <i>N</i>	14% <i>H</i>	28% <i>M</i>	34% <i>M</i>	9% <i>H</i>
Monitor text	94%	0% <i>H</i>	0% <i>H</i>	0% <i>H</i>	0% <i>H</i>	0% <i>H</i>	0% <i>H</i>	0% <i>H</i>	0% <i>H</i>	0% <i>H</i>	0% <i>H</i>	—
Environment												
Indoor general	97%	6% <i>H</i>	35% <i>H</i>	57% <i>N</i>	83% <i>N</i>	42% <i>H</i>	59% <i>N</i>	91% <i>N</i>	50% <i>H</i>	73% <i>N</i>	76% <i>N</i>	—
Indoor specific	94%	2% <i>H</i>	13% <i>H</i>	31% <i>H</i>	65% <i>N</i>	23% <i>H</i>	52% <i>N</i>	91% <i>N</i>	16% <i>H</i>	55% <i>N</i>	72% <i>N</i>	—
Outdoor general	100%	6% <i>H</i>	66% <i>H</i>	91% <i>N</i>	97% <i>N</i>	67% <i>H</i>	90% <i>N</i>	93% <i>N</i>	58% <i>H</i>	95% <i>N</i>	96% <i>N</i>	—
Outdoor specific	80%	4% <i>H</i>	24% <i>M</i>	42% <i>N</i>	72% <i>N</i>	16% <i>H</i>	23% <i>M</i>	80% <i>N</i>	20% <i>M</i>	48% <i>N</i>	58% <i>N</i>	—
Messy room	58%	0% <i>H</i>	17% <i>N</i>	23% <i>N</i>	27% <i>N</i>	30% <i>N</i>	21% <i>N</i>	57% <i>N</i>	18% <i>N</i>	51% <i>N</i>	46% <i>N</i>	—
Other												
Laundry	94%	17% <i>H</i>	26% <i>H</i>	48% <i>M</i>	48% <i>M</i>	21% <i>H</i>	40% <i>H</i>	57% <i>N</i>	31% <i>H</i>	28% <i>H</i>	44% <i>H</i>	39% <i>H</i>
Food	91%	22% <i>H</i>	37% <i>M</i>	48% <i>M</i>	41% <i>M</i>	37% <i>M</i>	42% <i>M</i>	76% <i>N</i>	33% <i>H</i>	20% <i>H</i>	50% <i>M</i>	57% <i>N</i>

Table 5.3: Recognition accuracy for different filters across different scenarios. Recognition accuracies are shown as percentages, while subscripts and colors indicate whether each filter is **effective (H)**, somewhat effective (M), or **not effective (N)** in preventing recognition, and asterices indicate significance: * is significant at $p < .05$, ** is significant at $p < .001$, and *** is significant at $p < .001$, after Bonferroni correction.

	<i>Masking</i>			<i>Blur</i>			<i>Pixel</i>			<i>Edge</i>			<i>Silhouette</i>
	P I S V	H i g h P I S V	M e d i u m P I S V	L o w P I S V	H i g h P I S V	M e d i u m P I S V	L o w P I S V	H i g h P I S V	M e d i u m P I S V	L o w P I S V	P I S V	P I S V	P I S V
Human													
Activity	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓
Age	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J
Dress	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓
Ethnicity	✓ X X X X	✓ X X X X	✓ X X X X	✓ X X X X	✓ X X X X	✓ X X X X	✓ X X X X	✓ X X X X	✓ X X X X	✓ X X X X	✓ X X X X	✓ X X X X	✓ X X X X
Expression	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓
Gender	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J
Hair	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J	✓ J J J J
<hr/>													
Document													
Document class	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓
Document type	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓
Document text	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓
<hr/>													
Monitor													
Monitor class	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J
Monitor application	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓	✓ X ✓ ✓ ✓
Monitor text	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓	✓ X ✓ ✓
<hr/>													
Environment													
Indoor general	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X
Indoor specific	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X
Outdoor general	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X
Outdoor specific	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X
Messy room	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X	✓ X X X
<hr/>													
Other													
Laundry	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J	✓ X J J J
Food	✓ X X J J	✓ X X J J	✓ X X J J	✓ X X J J	✓ X X J J	✓ X X J J	✓ X X J J	✓ X X J J	✓ X X J J	✓ X X J J	✓ X X J J	✓ X X J J	✓ X X J J

Table 5.4: Privacy and utility trade-offs. For each filter, a green checkmark or red cross indicates whether that filter 1) protects privacy (i.e. recognition accuracy < 50% and odds-ratio < 0.05) (P), 2) provides sufficient information (I), 3) creates a satisfactory image (S), and 4) creates a visually appealing image (V).

CHAPTER 6

Designing Novel Image Obfuscations

As the findings presented in chapter 5 suggest, most of the commonly used filters do not ensure adequate privacy protection; the ones that can protect privacy were not satisfactory to the viewers as they reduced the utility of the images. In the context of social media, preserving utility in filtered images is an important constraint that needs to be satisfied to drive the filters’ wide adoption in practice. This chapter details the steps that we followed to design novel image transforms and findings from a study to evaluate those transforms.

This study was done in collaboration with Yifang Li, Eman Hassan, Roberto Hoyle, David Crandall, and Apu Kapadia. Findings from this study were published as “Can Privacy Be Satisfying? On Improving Viewer Satisfaction for Privacy-Enhanced Photos Using Aesthetic Transforms” in CHI’19 [85].

6.1 Introduction

Chapter 5 presented a study in which we identified a set of obfuscations that can effectively obscure objects (or their properties) in a photo while minimizing the impact on the viewer’s overall “satisfaction.” However, the set of such useful obfuscations was relatively small; most obfuscations reduced the perceived “information sufficiency” or aesthetics to the point of negatively impacting people’s satisfaction in viewing the image. Since one of the primary motivations for sharing photos is to convey information and seek acceptance, appreciation, and validation from peers [130, 151], preserving the utility (satisfaction) of obfuscated images is important if obfuscation methods are to be widely accepted.

At a high level, obfuscation imposes a trade-off that is easy to understand: mild obfuscations may not negatively affect viewer satisfaction but may also not remove private image content effec-

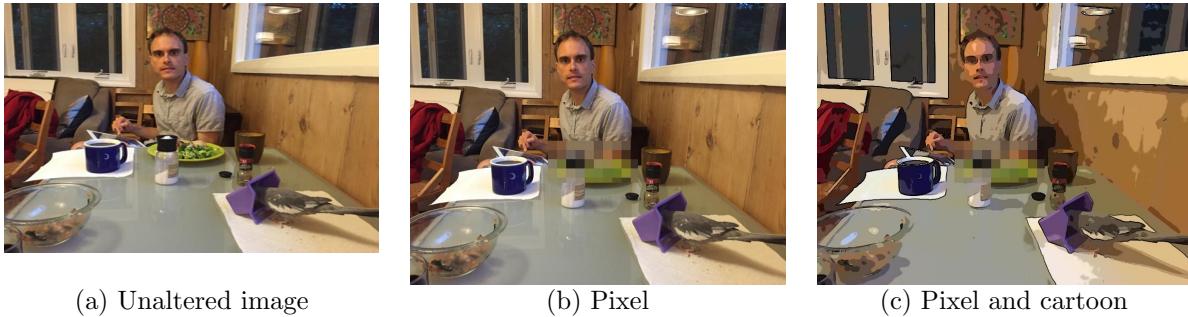


Figure 6.1: An example illustrating how obfuscation and beautification change the utility aspects of an image: (a) an image without any alteration, (b) the image after a pixel obfuscation, and (c) the image after a pixel obfuscation applied to the food plate and a cartoon beautification on the other parts of the image.

tively, while aggressive obfuscations may preserve privacy but cause obvious visual changes that reduce viewer satisfaction. Prior work identified three useful variables in measuring the viewer experience — information sufficiency, satisfaction, and aesthetics — and measured how obfuscations affect each of these variables in isolation [84, 125]. Those studies, however, do not examine the inter-relationships among these dependent variables. Further, in addition to direct effects, obfuscations might have cascading effects on these variables (i.e. affecting one variable through another). Understanding these relationships would greatly benefit in designing novel obfuscation methods that can improve privacy without adversely impacting viewers' experience.

Using data from the previous study, we conduct a path-model based analysis which suggests that the effects of the obfuscations on information sufficiency and visual aesthetics are much greater than the direct effects on satisfaction, but information sufficiency and visual aesthetics are significantly associated with satisfaction. This observation inspires our novel hypothesis that *it may be possible to compensate for the reduction in information sufficiency from obfuscations by increasing visual aesthetics, thus actually maintaining or improving overall satisfaction of the obfuscated image*. This approach is illustrated in Figure 6.1, where an object within the photo may be redacted with pixelation while the rest of the image is aesthetically ‘improved’ using an artistic transformation, resulting in satisfaction similar to the original image.

To test our hypothesis, we conduct a new online experiment with three obfuscation and three beautification transformations across a variety of photo types and scenarios. The experiment follows a between-subjects design that extends our previous experiment [84] by adding the beautification condition. Thus our design seeks to ascertain a causal relationship between the manipulation of aesthetics and the viewer’s overall satisfaction with various obfuscations. From the photo aesthetics literature [149], we know that colors and tones play an important role in aesthetics: pure and high saturation colors tend to be more appealing to viewers than dull colors, for example [47]. We pick three particular beautifications to represent different levels of abstraction: (1) a low-level abstraction using color correction [62] to produce an effect similar to highly popular Instagram filters [131], (2) a ‘cartooning’ effect similar to a watercolor painting that moderately changes the appearance of the original image, and (3) a deep learning-based algorithm to render the photo in an bright, colorful style, inspired by the popular Prisma app [107], that produces a highly abstract and unrealistic version of the image. We refer to these three beautification transformations as ‘colors’, ‘cartoons’, and ‘abstract’, respectively.

Our results verify interactions among information content, aesthetics, and satisfaction, confirming that it is worthwhile to investigate whether satisfaction can be increased by increasing the other two variables. Although the gain in satisfaction was not statistically significant for our sample data using off-the-shelf artistic transforms, we hope our findings will inspire designing new transforms taking into account the negative effects of privacy obfuscations.

6.2 Method

In earlier work, we studied the privacy-protecting and utility-preserving qualities of four obfuscation methods (i.e., image filters) [84]. There we applied these filters on people and other objects to obfuscate properties or attributes (such as the age of a person, the organization of a room) that were identified as privacy sensitive in prior work. We experimented with 20 attributes, and

analyzed how effective each of the filters was in obscuring these attributes and how they affected the utility variables (i.e., information content, aesthetics, and viewers' satisfaction). In this work We conducted additional analysis of that data using path models to study the inter-dependencies of the utility variables. The next two subsections describe the procedure and results of this analysis. We then provide details of our new experiment, which was inspired by the results of the path model analysis.

6.2.1 Path Model Analysis

We constructed separate path models using data from our previous experiment [84] for each of the 20 attributes (e.g., activity, gender, document class, document type). In these path models the exogenous variable was obfuscation type (such as blur and pixel) and the endogenous variables were information sufficiency, photo aesthetics, and photo satisfaction. We excluded data about identification accuracy and confidence from our model since we focus on the utility variables. For each attribute, we began with the initial model shown in Figure 6.2, and then trimmed insignificant effects.

In this graph representation the vertices represent variables, and arrows represent relationships between the variables. Further, the blue rectangular vertices are the exogenous variables (e.g., different obfuscations), and the orange ellipses are the dependent variables measured (e.g., information content). The directional edges in this graph express causal relationships, where changing the variable denoted by the starting vertex of an edge ‘causes’ a change in another variable denoted by the finishing vertex of the same edge (e.g., changing the obfuscation method causes a change in ‘satisfaction’). This model also captures indirect causal effects, such as obfuscation methods’ effects on ‘satisfaction’ through ‘information content’. The causal effects between the endogenous variables are speculative, but we describe the rationale for this particular arrangement of the vertices and the directions of the edges, e.g., why we think ‘information content’ causally affects ‘satisfaction’

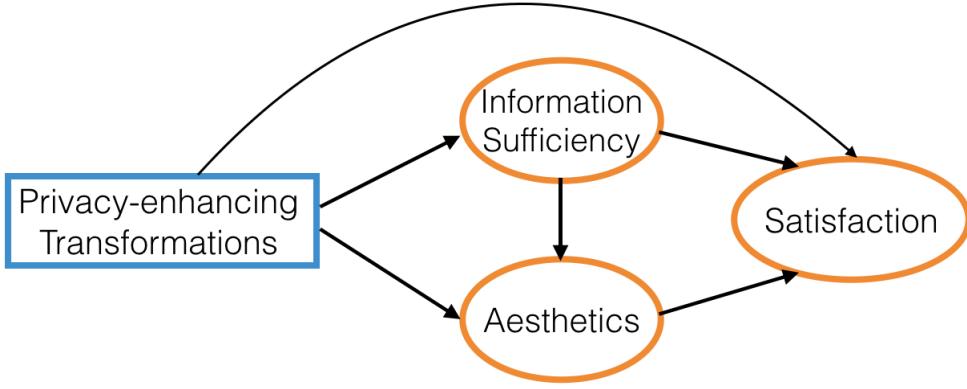


Figure 6.2: Initial path model.

and not the other way around.

An important motivation for people to use online social networks is to gather information, such as by observing other people's photos [206]. This means that for high satisfaction, viewers need to be able to see important content ('sufficient information') in the photo. Aesthetics also contributes to satisfaction; in fact, users often edit their photos before sharing to improve aesthetics and to help control the impressions conveyed to others [186]. Our initial path model thus assumes there are causal relationships (and our experiment seeks to test such causality) from information sufficiency and aesthetics to satisfaction. For example, increasing aesthetics or information sufficiency may improve satisfaction when viewing the photo. From a photo composition perspective, we expect that displaying sufficient information improves photo aesthetics. Finally, based on previous work that shows that obfuscations affect information content sufficiency, photo aesthetics, and satisfaction in most scenarios, our initial model includes causal arrows from transformations to each dependent variable [84].

6.2.2 Path Model Results

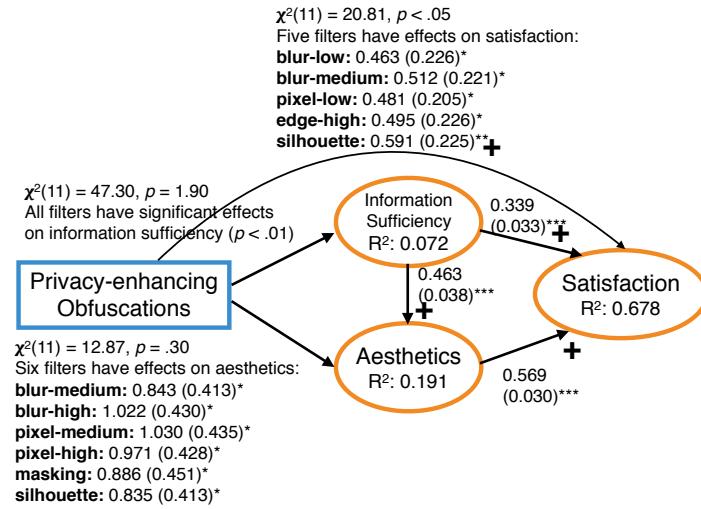
As expected, our findings generally indicate that obfuscations have negative effects on information sufficiency, while their effects on aesthetics vary based on attribute types. For example, in the

laundry scenario (Fig. 6.3a), applying obfuscation has no effect on image aesthetics ($\chi^2(11) = 12.87$, $p = 0.30$). A possible explanation is that laundry is typically not an appealing or important visual element, so obscuring it does not affect the aesthetics of the overall photo. Additionally, in half of the scenarios (age, document type and text, dress, ethnicity, expression, food, hair, indoor general and specific, and messy room), there is no direct effect on photo satisfaction by different types of obfuscations, although there are indirect effects mediated by information sufficiency and aesthetics.

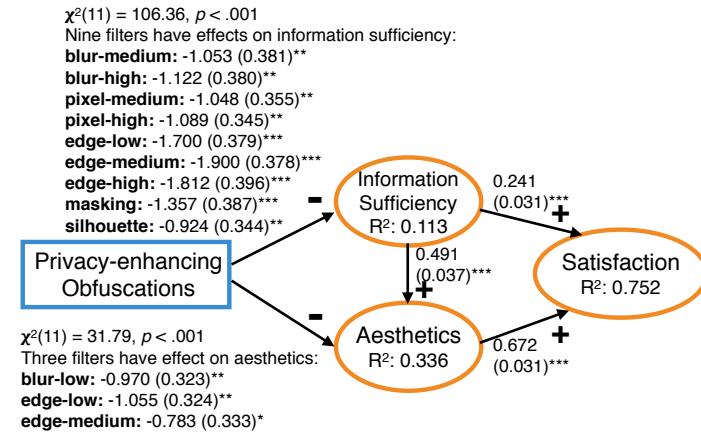
For example, consider the path model for *dress* (Figure 6.3b), which has a good model fit ($\chi^2(11) = 13.02$, $p = 0.29$, $CFI = 0.998$, $TLI = 0.994$, $RMSEA = 0.018$). Overall, there is a difference in information sufficiency between different transformation conditions ($\chi^2(11) = 106.36$, $p < 0.001$) compared to the baseline condition; most of the transformations (blur-medium, blur-high, pixel-medium, pixel-high, edge-low, edge-medium, edge-high, masking, and silhouette) decrease information sufficiency (all $p < 0.01$) while blur-low and pixel-low do not have any effect. On the other hand, obfuscations also have a generally negative effect on aesthetics ($\chi^2(11) = 31.79$, $p < 0.001$). Photos on which blur-low, edge-low, and edge-medium have been applied have lower aesthetics compared with the original photos (all $p < 0.05$). Meanwhile, information sufficiency appears to have a significant effect on aesthetics ($p < 0.001$), with a one-point difference in information sufficiency associated with a 0.491-point difference in aesthetics ($SE = 0.037$). Furthermore, aesthetics appears to positively affect satisfaction ($p < 0.001$), and information sufficiency also appears to have a direct effect on satisfaction ($p < 0.001$).

More generally, in all scenarios, controlling for manipulations, information sufficiency has a highly significant positive association with aesthetics (all $p < 0.001$). Additionally, aesthetics (all $p < 0.001$) and information sufficiency (all $p < 0.001$) have a direct positive association with satisfaction. These results indicate that increasing either information sufficiency or aesthetics may boost image satisfaction, and beautification on the remaining (non-obfuscated) part of the image could make

Figure 6.3: Example path model diagrams.



(a) Path model for *Laundry*



(b) Path model for *Dress*

Privacy enhancing obfuscations	Beautification transformations
Masking	Abstract
Pixelation	Cartoon
Edge	Color

Table 6.1: Obfuscations and transformations used in this study. Each obfuscation was combined with each transformation, resulting in nine conditions. In addition, we included 3 obfuscation-only conditions, as well as a condition with the original, unaltered image, totaling 13 experimental conditions.

up for the viewers' satisfaction lost through obfuscation. To test this causal effect, we conducted a new online experiment; the design and methodology of this experiment is described in the following sections.

6.2.3 Experimental Design

In this experiment, we presented participants with photos that had been manipulated using various combinations of privacy-enhancing obfuscation and beautification transformations, and collected their ratings on utility variables. The privacy-enhancing obfuscations were applied on specific regions of a photo (to obscure attributes of people/objects) and the beautification transforms were applied on the rest of the photo. With 3 obfuscations and 3 beautification transforms, our study had 13 between-subjects experimental conditions (3 obfuscations + 3 obfuscations \times 3 beautifications + 1 unfiltered) (see Table 6.1). The baseline (i.e. unfiltered) condition included images without any alteration. The other conditions had only an obfuscation or an obfuscation combined with a beautification. Participants were randomly assigned to one of these conditions (between subjects), but each participant viewed images for all six object attributes (described below). Similar to our prior study [84], each participant viewed an image for each attribute and then answered five questions corresponding to the five dependent variables that we measured, as described below.

6.2.4 Participants

For our previous study [84], the number of participants per condition was calculated using a power analysis based on data from a pilot study. We planned a similar number of conditions and analysis

Attribute	Question
Document class	What is the object inside the green rectangle?
Dress	What type of clothing is the person inside the green rectangle wearing?
Gender	What is the gender of the person inside the green rectangle?
Laundry	What is the object inside the green rectangle?
Computer app.	What application is displayed on the computer monitor inside the green rectangle?
Monitor text	What is the text inside the green rectangle?

Table 6.2: The six attributes and corresponding detection questions used in the survey.

for this new study, hence we used the same number of participants (48) for each condition. With thirteen conditions, we needed at least $13 * 48 = 624$ participants in total. We advertised our experiment on Amazon Mechanical Turk¹ and hosted it on Qualtrics², restricting participation to MTurk workers with a high reputation (above 95% approval rating on at least 1000 completed HITs) to ensure data quality [137]. We also required workers to be at least 18 years old and living in the United States for at least five years to help control the cultural variability [106]. We included three attention check questions to maintain data quality [128]. After removing the responses from participants who provided wrong answers for one or more attention checks, we were left with 653 responses (out of a total of 780) that we used for analysis. Each participant was paid \$1.50, whether or not we used their response. The study was approved by Indiana University’s ethics board.

6.2.5 Selecting Attributes

From the set of twenty privacy-sensitive attributes used in our earlier experiment [84], we selected six to include in this study (see Table 6.2). We chose these six attributes to balance the size of the private image regions, since the sizes of obfuscated regions may otherwise vary dramatically depending on the size of the object to be obfuscated and/or the attribute itself. For example, we did not include any scenarios where the whole image needed to be obfuscated (e.g., hiding whether a photo was taken indoors or outdoors), since we wanted to study our hypothesis in the context of object obfuscations.

¹<https://www.mturk.com>

²<https://www.qualtrics.com>

6.2.6 Image dataset

We used the same image set we previously used in [84], which allowed us to isolate and measure the effects of beautifications on the filtered images in this experiment. The dataset contains sets of five images for each attribute, all collected from online sources. Care was taken to ensure that all images in each set were consistent with each other in terms of the number of objects and people, the shapes and sizes of these objects, the overall image quality and brightness, and the effort required to infer a certain attribute.

6.2.7 Privacy-enhancing Transformations and Artistic Transformations

We identified three main obfuscations: masking, pixelation, and edge. Previous work applied each of these transformations with three strength levels (high, medium, and low) and found that the ‘high’ level was most effective at obscuring sensitive attributes [84], so we use only that level here.

We chose three different beautifications that abstract scene content to different degrees. Our most conservative transformation, which we call *No-abstraction* or ‘Colors’, applies the color correction technique of Finlayson *et al.* [62], which modifies colors but does not affect the semantic content of the image. *Mid-abstraction* or ‘Cartoons’ applies a simple technique for “cartooning” the image, by applying bilateral filter-based blurring (Tomasi and Manduchi [205]), detecting edges from image gradients and highlighting them in black, and performing luminance quantization to 8 levels. This beautification abstracts some image content, since the blurring reduces resolution and the luminance quantization and added edges create an artistic effect. Finally, *Max-abstraction* or ‘Abstract’ applies deep-learning based artistic style transfer [223] for Henri Matisse’s famous painting *Woman with a hat*. Using artistic transforms that abstract photo content to compensate for lost information (due to the application of privacy obfuscations) might seem counter intuitive; we hypothesize that, since the abstraction happens at the global level, local information loss due to obfuscations may be less noticeable. Further, the abstracted form of the photos may help viewers

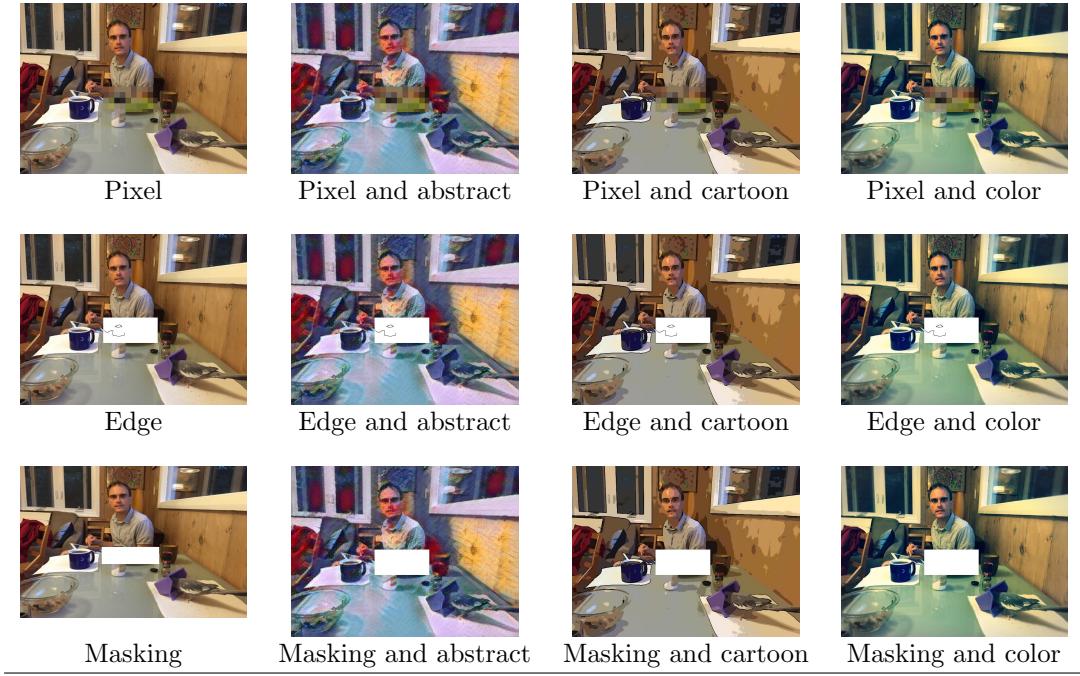


Table 6.3: Results of applying different obfuscations and beautifications.

absorb the high level story of a photo more easily, thus creating a sense of complete information.

Our experiment used 13 versions of each image as shown in Table 6.3: unaltered, obscured (3 versions), and obscured and beautified (3x3 versions), resulting in the thirteen experimental conditions. For privacy-enhancing transforms, we used the same transformation size, position, and other parameters reported in [84]. For the obscured and beautified versions, we first applied the obfuscations on the specific image regions, and then the artistic transform to the rest of the image using one of the three beautifications. The obscured areas were not beautified to hold the degree of privacy constant when comparing the obscured version with the obscured and beautified version; otherwise a higher satisfaction score could be attributed to lower privacy through first obscuring and then beautifying a sensitive object.

6.2.8 Measurements

For each attribute, we asked five questions from two perspectives: obfuscation effectiveness and utility to the viewer. We note that all these questions and response options were adapted from our previous study [84].

1. **Identification.** Participants first saw an image with a green bounding box overlaid on an object of interest. They were asked to identify the object in the box by answering a multiple-choice question, “What is the object (or property of the object) depicted in the image?” The specific questions were slightly different based on the attribute, as shown in Table 6.2. For this question, we provided a list of options (including “Cannot tell”) to select from as an answer. The green bounding boxes surrounding the objects/attributes of interest were shown only in this question and not for the following ones.
2. **Identification Confidence.** Participants answered “How confident do you feel that you correctly answered the previous question?” on a seven-point Likert scale from 1 ‘Completely unconfident’ to 7 ‘Completely confident’ [161].
3. **Information Content Sufficiency.** We asked participants to rate their agreement with “The photo provides sufficient information,” on a 7-point Likert from ‘Strongly disagree’ to ‘Strongly agree.’ This item was adapted from the ‘information quality scale’ [179], which measures “the satisfaction of users who directly interact with the computer for a specific application.” Our item loads onto the “content” factor and is strongly correlated with “is the system successful?” [179]
4. **Visual Aesthetics.** To measure photo aesthetics, we used “This photo looks visually appealing” from the image appeal scale [46], again on a 7-point Likert scale.
5. **Satisfaction.** Similarly, “The photo is satisfying” was adapted from the image appeal scale [46], which has also been used when measuring satisfaction of face and body obfuscation [84].

cation [125]. This item measures participants' overall satisfaction with the photo and again was rated on a 7-point Likert scale.

6.2.9 Procedure

The experiment flowed as follows:

1. Consent form detailing the experiment, estimated time to finish, and compensation.
2. Questions about social media usage and frequency of image sharing activities, along with demographics.
3. Instructions on how to respond to the survey questions with a sample image and questions.
4. Six blocks of questions corresponding to the six attributes, in random order. Each block showed the five questions corresponding to the five measurements for each attribute. One of the five photos for each attribute was randomly selected to be presented to the participant with the assigned condition ('unaltered', 'obfuscated', or 'obfuscated plus beautified').

6.2.10 Data Analysis Procedure

We used non-parametric versions for all of our statistical tests as our data do not meet the assumptions of parametric tests, such as normality and equal variance of errors. For each dependent variable (information content, visual aesthetics, satisfaction), we first conducted an overall Kruskal-Wallis test across all conditions to see if there was any significant difference in the measured variables among the conditions. We followed this with a Dunn's post hoc test with Bonferroni correction, where we compared between specific pairs. For each dependent variable, we selected the pairs to compare as follows: for each of the three obfuscation conditions (masking, pixel, edge) was compared with the three corresponding *obfuscation plus beautification* conditions. Therefore, for each of the three obfuscations, we had three pairwise tests, for a total of nine. This set of pairwise

tests allowed us to study whether combining beautification transforms with privacy obfuscations increases the utility of photos. Next, we conducted additional pairwise tests to see how combinations of privacy obfuscations and beautification transforms preserve utility when compared with the original (i.e. unaltered) photos. To do this, for each of the three obfuscations, we picked one beautification transform that performed best (i.e. highest mean value of the measured variable) when combined with it, yielding three *obfuscation plus beautification* conditions. Then these *obfuscation plus beautification* conditions were compared with the *unfiltered* condition. This resulted in three additional comparisons, or twelve in total. We present results of these pairwise tests in the supplementary document, where, in addition to the test statistics, we report the Pearson's product moment correlation (r) [43].

As an example of the process, for the *dress* attribute and the information content dependent variable, we first conducted an overall Kruskal-Wallis test for any difference in information content across the experimental conditions. If the p-value was not significant, we did not conduct any follow-up. If the p-value was significant ($p < 0.05$), then there were significant differences involving at least two different conditions. To find the pairs of conditions having differences, we followed up with Dunn's post hoc test for pairs of *only obfuscation* and *obfuscation plus beautification*. For example, for the *masking* obfuscation, we compared the *masking* condition with each of *masking + abstract*, *masking + cartoon*, and *masking + color* applied on the *dress* attribute. Also, if for example *masking + abstract* retained more information among these three *obfuscation plus beautification* conditions, we compared it to the *unfiltered* condition for the same measured variable (i.e., information content). This setting allowed us to test the effects of beautifications on obfuscated images, and also study the behavior of obfuscation-beautification combinations compared with the *unfiltered* condition.

6.3 Findings

We now present the results of our experiment.

6.3.1 Demographic Characteristics of the Participants

Out of 653 participants, 436 (66.7%) identified themselves as male and 216 (33%) as female. Our participants were typically under 49 years of age, with 351 (53.7%) between 18 and 29 years, 250 (38.3%) between 30 and 49 years, 54 (6.7%) between 50 and 64 years, and eight (1.2%) participants 65 years or older. Three hundred and thirty five (51.3%) participants were white, 152 (23.2%) were Asian, and 43 (6.5%) were black or African American. For the highest level of education, 320 (49%) participants reported an undergraduate degree, 172 (26.3%) high school, 145 (22.2%) a Master's degree, and 16 (2.4%) a professional degree. All participants reported having at least one social network account, while 512 (78.3%) reported sharing photos online with frequency ranging from several times a day to a few times a week, and only 25 (3%) participants reported never posting photos online.

6.3.2 Effects of Transformations on Information Content

For all attributes, perceived information content was the highest for the *unfiltered* condition (Table 6.4). The *abstract* transform, when combined with privacy obfuscations resulted in the lowest information content for most of the attributes (Table 6.4). Surprisingly, the *color* transform, which alters the image content the least, reduced more information than the *cartoon* transform, which, when combined with *edge* and *pixelation* privacy obfuscations, actually increased perceived information content for most of the attributes (Table 6.4). We conducted an overall Kruskal-Wallis test and detected significant differences in perceived information content among different *obfuscated*, *obfuscated plus beautified*, and *unfiltered* conditions (*document*: $\chi^2(11) = 54.75$, *dress*: $\chi^2(11) = 55.55$, *gender*: $\chi^2(11) = 57.55$, *computer application*: $\chi^2(11) = 81.00$, *monitor text*: $\chi^2(11) = 45.26$,

laundry: $\chi^2(11) = 39.07$, all $p < 0.01$.

Next, we conducted Dunn's post-hoc pairwise tests with Bonferroni correction to detect any significant differences in information content (see supplementary document). For all attributes, pairwise Dunn's tests comparing the *only obfuscation* and *obfuscation plus beautification* conditions revealed no significant difference in information, meaning that combining *beautification* with *obfuscation* does not reduce any more information. When compared with the *unfiltered* condition, we found that for *gender*, *computer application*, and *monitor text*, all *obfuscation plus beautification* transforms resulted in significant reduction in information content with medium to high effect sizes ($.45 \leq r \geq .75$, all $p < .01$). For *document* and *dress*, except for *edge + cartoon* and *pixelation + cartoon* respectively, all other *obfuscation plus beautification* transforms significantly reduce information content ($.43 \leq r \geq .6$, all $p < .05$). Finally, for *laundry*, only the *pixelation + cartoon* transform results in reduction in information with medium effect size ($r = .42$, $p < .05$).

Overall, despite being a source of additional abstraction, the beautification transforms do not cause any significant additional reduction in information content to an obscured image.

6.3.3 Effects of Transformations on Visual Aesthetics

Overall Kruskal-Wallis tests indicated that there are significant differences in perceived image aesthetics across conditions for all attributes except *monitor text* (*gender*: $\chi^2(11) = 21.26$, *dress*: $\chi^2(11) = 37.75$, *document*: $\chi^2(11) = 21.54$, *computer application*: $\chi^2(11) = 23.88$, *laundry*: $\chi^2(11) = 24.36$, $p < 0.01$ for *dress*, $p < 0.05$ for other attributes). For the *document* and *dress* attributes, the *unfiltered* condition has the highest scores for visual aesthetics, and combining aesthetic transforms reduced scores compared to applying only privacy obfuscations (Table 6.5). For *document*, the reductions were not significant for any beautification transform (all $p > .05$) but for *dress*, combining *cartoon* with *masking* significantly lowered visual aesthetics ($z = 3.02$, $r = .42$, $p < .05$) compared to the condition when only the *masking* obfuscation was applied. On the other

Condition	Document		Dress		Gender		Laundry		Computer app		Monitor text	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Unfiltered	5.34	1.24	5.56	1.36	5.48	1.16	5.18	1.37	5.42	1.49	5.26	1.55
Edge	4.76	1.55	4.13	2.30	3.41	2.07	3.37	1.96	3.17	2.08	3.43	2.23
Edge + Abstract	4.04	1.87	3.49	2.00	3.40	2.13	3.49	1.94	3.00	2.02	3.36	2.13
Edge + Cartoon	4.64	1.64	4.22	1.97	3.94	1.90	4.24	1.64	3.74	1.75	3.86	1.74
Edge + Color	4.18	1.99	4.12	2.00	3.35	2.11	3.37	1.95	3.47	2.27	3.55	2.20
Pixelation	3.90	2.04	4.84	1.70	3.90	2.04	3.82	1.86	3.78	1.81	3.57	2.33
Pixelation + Abstract	3.76	1.62	4.25	1.75	3.57	1.70	3.96	1.75	3.90	1.88	3.33	1.91
Pixelation + Cartoon	4.08	1.61	4.98	1.52	3.67	1.93	4.08	1.61	3.92	1.70	3.17	1.72
Pixelation + Color	3.66	1.91	4.26	1.81	3.64	1.88	3.60	1.82	3.43	1.80	3.19	2.15
Masking	3.98	1.97	4.24	2.11	3.58	2.17	3.90	1.94	3.34	2.07	3.42	2.01
Masking + Abstract	3.38	1.89	3.27	1.76	2.88	1.62	4.06	1.83	2.77	1.77	3.17	1.86
Masking + Cartoon	3.74	1.88	4.43	1.91	3.61	1.95	4.13	2.05	3.22	1.98	3.41	2.01
Masking + Color	3.38	1.74	4.02	1.85	2.98	1.81	3.80	1.78	2.38	1.60	3.00	1.96

Table 6.4: Means and standard deviations of information content scores for different attributes.

hand, for *laundry*, *edge* obfuscation combined with the *cartoon* transform produced significantly more visually appealing photos compared to both when only *edge* obfuscation was used ($z = 3.01$, $r = .42$, $p = .03$) and the *unfiltered* condition ($z = 3.03$, $r = .43$, $p < .05$). For *gender* and *computer app*, no *obfuscation plus beautification* transform significantly increased aesthetics over *only obfuscation* conditions (see supplementary material).

Overall, except for the *cartoon* transform (in one case), the beautification transforms did not significantly increase the visual aesthetics of obscured photos.

6.3.4 Effects of Transformations on Viewers' Satisfaction

Except for the *computer application* and *monitor text* attributes, photo satisfaction had the highest scores in the *unfiltered* condition for all other attributes (Table 6.6). Kruskal-Wallis tests across all conditions detected significant differences in satisfaction scores for *document* ($\chi^2(11) = 22.38$, $p < .05$), *dress* ($\chi^2(11) = 47.1$, $p < .0001$), and *computer app* ($\chi^2(11) = 25.96$, $p < .01$). For the other three attributes, none of the conditions had significantly different satisfaction scores than others. Hence we conducted pairwise tests only for *document*, *dress*, and *computer app* attributes (see supplementary material). We did not find any statistically significant increase in satisfaction when comparing *only obfuscation* with *obfuscation plus beautification* (see supplementary material). Finally, comparing the *obfuscation plus beautification* conditions with the *unfiltered* condition, we found that for *dress*, *masking + cartoon* significantly lowered satisfaction ($z = 3.6$, $r = 0.51$, $p < .001$). All other results were non-significant ($p > 0.05$).

Overall, we did not find statistically significant evidence that our selected artistic transforms increase viewers' satisfaction compared to obscured photos.

Condition	Document		Dress		Gender		Laundry		Computer app		Monitor text	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Unfiltered	4.92	1.65	5.00	1.29	4.24	1.59	3.62	1.99	3.48	1.55	3.50	1.64
Edge	4.54	1.62	4.20	1.75	3.87	1.78	3.65	1.99	3.54	2.03	3.67	1.93
Edge + Abstract	3.80	1.66	3.38	1.57	3.38	1.93	3.76	1.79	3.20	1.82	3.47	1.67
Edge + Cartoon	4.62	1.71	4.38	1.81	4.34	1.79	4.72	1.75	4.08	1.84	4.20	1.74
Edge + Color	4.04	1.98	3.80	1.77	3.47	1.82	3.61	1.99	3.43	1.88	3.53	1.97
Pixelation	4.57	1.70	4.61	1.59	4.31	1.59	3.92	1.65	3.71	1.80	3.90	1.98
Pixelation + Abstract	4.53	1.47	4.33	1.65	4.27	1.63	4.35	1.60	4.37	1.62	4.16	1.60
Pixelation + Cartoon	4.31	1.68	4.44	1.77	4.06	1.62	3.96	1.71	3.58	1.77	3.73	1.77
Pixelation + Color	4.26	1.93	4.09	1.79	3.92	1.80	3.57	1.86	3.26	1.77	3.51	1.72
Masking	4.76	1.92	4.96	1.73	4.26	1.87	4.34	1.94	3.64	1.99	3.78	1.93
Masking + Abstract	4.46	1.74	4.23	1.64	4.50	1.68	4.42	1.53	3.81	1.66	3.83	1.58
Masking + Cartoon	3.93	1.83	4.00	1.65	3.78	1.63	3.93	1.80	3.22	1.90	3.61	1.75
Masking + Color	4.48	1.76	4.48	1.47	4.06	1.65	3.92	1.76	3.06	1.63	3.34	1.56

Table 6.5: Means and standard deviations of visual aesthetics scores for different attributes.

Condition	Document	Dress	Gender	Laundry	Computer app	Monitor	text
	mean	std	mean	std	mean	std	mean
Unfiltered	4.84	1.42	5.32	1.45	4.62	1.46	4.18
Edge	4.43	1.77	4.20	1.98	3.69	1.97	3.87
Edge + Abstract	3.80	1.82	3.31	1.73	3.36	1.91	3.67
Edge + Cartoon	4.74	1.60	4.42	1.89	4.14	1.68	4.52
Edge + Color	4.06	1.92	3.67	1.95	3.51	1.95	3.65
Pixelation	4.45	1.79	4.71	1.65	4.24	1.74	4.29
Pixelation + Abstract	4.27	1.66	4.25	1.61	4.02	1.67	4.20
Pixelation + Cartoon	4.31	1.49	4.65	1.54	3.94	1.73	3.75
Pixelation + Color	3.92	1.81	4.15	1.81	3.89	2.01	3.64
Masking	4.14	1.82	4.58	1.91	4.06	1.85	4.36
Masking + Abstract	4.02	1.74	3.83	1.59	3.96	1.69	4.19
Masking + Cartoon	3.70	1.88	4.02	1.95	3.70	1.92	4.04
Masking + Color	3.92	1.93	4.00	1.65	3.68	1.78	3.70

Table 6.6: Means and standard deviations of photo satisfaction scores for different attributes.

6.4 Discussion

Our hypothesis was that by applying beautification techniques to an image in which privacy-sensitive content has been obfuscated, we can increase both perceived information content (possibly by providing a high-level story) and visual aesthetics, and thus recover some or all of the viewer satisfaction that would otherwise have been lost to the privacy transform. Our results show that the beautifications we experimented with did *not* significantly increase satisfaction. Certain combinations of obfuscation and beautification transforms (e.g., when using the *cartoon* transform), however, appeared to increase some or all of the three dependent variables (information content, visual aesthetics, and satisfaction). These combinations could be studied with more statistical power in the future, or with modifications that attempt to increase aesthetics and satisfaction. It is interesting to see that the *cartoon* transform boosted information sufficiency despite being a form of abstraction. This supports our speculation that presenting photo content at a high level might increase overall information absorption. Also it may be that viewers found the ‘beautified’ versions more interesting and derived more information from the transformed photo. For example, an ordinary object may appear more interesting following the cartoon transformation.

We found that the *abstract* transform appeared to increase aesthetics in some cases, but lowered information content without increasing viewers’ satisfaction; we expected a greater increase in perceived visual aesthetics since this is the most artistic transform among the three. One possible explanation is that the reduction in information content by the *abstract* transform might negatively affect the other two variables, since our results from the path model analysis show that information content is associated with both of those variables. Finally, we found the *color* transform did not increase any of the measured variables. We expected *color* to have a lesser effect on both lowering information content and increasing visual aesthetics compared to the other two transforms. It might be the case that the negative effect of the loss of information on visual aesthetics and satisfaction was greater than the increase, if any, in the latter two variables.

Although we did not have sufficient statistical power to ascertain the difference in satisfaction between the obfuscated and beautified conditions, our findings still suggest the validity of an approach where a combination could increase satisfaction, with the cartoon filter being the most promising. Overall, we believe future work should explore other possible beautification transforms to study the novel privacy vs. satisfaction trade-off. It may be particularly promising to study obfuscating transforms that are themselves aesthetically pleasing or ‘fun’ instead of beautifying the rest of the image – as people grow accustomed to filters and effects (such as ‘stickers’) in photo-sharing applications, it will be increasingly acceptable to apply such obfuscations and transforms in general. By understanding and quantifying the effects of obfuscation on privacy and satisfaction, as well as the effects of beautification on satisfaction, we may be able to design the ‘correct’ combination of transformations for sensitive and non-sensitive image regions in order to both improve privacy and retain (or improve) satisfaction for the viewer. Indeed, improving privacy could be ‘fun’ too, both for the person transforming the photo and the viewer.

6.4.1 Limitations

We note several limitations of our study, which could be addressed in future work. We purposely restricted our pool of MTurk participants to users in the United States of at least 18 years of age to control for cultural differences. Although MTurk participants resemble US population fairly well and better than other web panels [171], our findings may not generalize for other age groups. Further, photo sharing behaviors as well as perceptions of privacy and aesthetics differ across cultures, and explicitly studying these differences in the context of beautification and obfuscation transformations would be interesting for future work. Moreover, we used the same pool of photos as past work to allow for direct comparison with published results, but these photos were collected from web sources. Participant views of aesthetics and satisfaction on these images may not reflect how they would feel about transformations applied to their own images. Follow-up studies could

request users to subject their own photos to transformations, and compare outcomes on those photos versus the web images we consider here. Our selections of obfuscation and beautification transformations were made based on past work, and they were designed precisely for the same purposes as ours – to obfuscate objects and increase photo aesthetics. There are many other possible combinations of such transformations, and studying a larger set may reveal techniques that are more effective at balancing privacy, aesthetics, and satisfaction. Finally, we did not consider other obfuscation techniques (such as Snapchat filters and Apple Memoji) that can add or replace information instead of just obscuring (e.g., a smiley face replacing the original emotion of a person). While the popularity of these features indicates their effectiveness in retaining and/or increasing viewers’ satisfaction, it would be interesting to study their effectiveness in protecting privacy.

6.5 Conclusions

We explored the novel question of whether a viewer’s satisfaction of a photo with obfuscated elements can be improved. While one might expect there to be a strict privacy-satisfaction trade-off, where applying obfuscations to improve privacy degrades the viewing experience, we hypothesize that ‘beautification’ transforms can be applied to the *rest* of the image to compensate for or counteract the loss in satisfaction, in order to create an image that *both* preserves privacy and viewer satisfaction.

As a first step, we experimented with three off-the-shelf beautification transforms and extended prior work on obfuscation transforms to evaluate combinations of obfuscation and beautification. While we did not find statistically significant support for our hypothesis that these transforms boost viewers’ satisfaction, we hope the gain in information content and visual aesthetics will inspire the exploration of new transforms that take into account the negative effects of privacy obfuscations, as well as obfuscating transforms that are themselves aesthetically pleasing (e.g., a sticker obfuscating a face but also making the image more fun to look at). We believe this line of work is particularly

salient with the popularity of photo sharing and adding photo effects and stickers, and hope it inspires further exploration of how such transforms can be used not only for entertainment but to simultaneously afford more privacy.

CHAPTER 7

Individual Differences and Photo-sharing Behaviors

As Chapters 1 and 3 pointed out, sharing memes poses severe privacy threats to the people appearing in those photos. But it is difficult to provide any technical means through which the photo subjects may exert any control over the dissemination of memes featuring them. An alternative approach to reduce such privacy violations may be to raise awareness among social media users, who create and or disseminate memes, about how such activities may harm the photo subjects and stimulate privacy-respecting and pro-social behavior using behavioral interventions (e.g., privacy nudges). Existing research in this direction has been scarce. Recently, Amon *et al.* conducted an experiment where the participants were primed to consider the privacy of people in memes [16]. Surprisingly, they found that participants who were primed shared more memes compared to a control group. Much other research related to security and privacy decision making reported lower effectiveness of ‘generic’ (i.e., designed to alter the behavior of an ‘average’ person) interventions than was expected; and researchers have advocated for personalized interventions [109] which were found to be more effective in several cases [19, 127, 142, 217]. But a comprehensive understanding of the decision-making process regarding photo-sharing, and the underlying factors that influence this process is a prerequisite of designing effective behavioral interventions. In this chapter, we take a step back and attempt to understand how *individual differences in the usage of humor* affects people’s meme sharing decisions and their reactions to privacy nudges designed to discourage them from sharing memes. Independent of assisting in developing effective and personalized interventions, this understanding would advance our knowledge of the human decision-making process.

This work was done in collaboration with Bennett I. Bertenthal, Kurt Hugenberg, and Apu Kapadia and will be published as “Your Photo is so Funny that I don’t Mind Violating Your Privacy by Sharing it: Effects of Individual Humor Styles on Online Photo-sharing Behaviors” in

7.1 Introduction

Image macros or photo-memes (referred to by ‘memes’ for the rest of the chapter) are made out of photos that often contain additional texts that suggest interpretations of those photos which are unrelated to the original context of those photos. Circulations of memes goes outside of the ‘imagined audience’ [10, 120] and may lead to ‘context collapse’ [29, 30, 150]. There have been many occasions where people appearing in memes were maligned or embarrassed in front of a large audience, leading to psychological distress and disruption in their professional and personal lives [2, 21, 49]. Yasmineen *et al.* reported some of the manual strategies undergraduate students adopt to avoid being photographed by others and then becoming portrayed in memes as it would adversely affect their reputation and future prospectus of employment [168], but there is a lack of technological solutions to this problem.

Recently, Amon *et al.* published surprising findings – nudges designed to reduce the sharing of photo-memes *amplified the unintended behavior*, i.e., participants who were primed demonstrated *higher* sharing likelihood compared to the control group [16]. This suggests that a deeper understanding of what personal factors drive people’s meme-sharing behaviors so that for an individual person such behaviors can be predicted based on relevant personality traits and personalized interventions can be applied.

To step forward in this direction, in this paper, we report the findings from a study we conducted to understand whether the individual humor type (i.e., how one uses humor to entertain the self or advance social relationships [133]) i) influences sharing photos of other people on social media and, ii) dictates how one would react to behavioral interventions designed to encourage privacy-respecting behaviors. In an online study, we asked the participants to indicate the likelihood of them sharing photo-memes on social media. In addition to the control condition, participants were

primed by i) instructing to imagine themselves as the subjects in the memes and ii) explicit warnings about potential privacy violations. In each condition, a time-delay intervention was employed such that participants had to view the memes for eight seconds before they could indicate their likelihoods to share the memes. This was done to ensure that participants had enough time to examine the memes carefully and think along the line of the intervention (when present) rather than acting impulsively [9, 144]. We also collected data about participants' history of sharing other people's (both familiar and stranger) photos in real-life and humor styles using the Humor Style Questionnaire (HSQ) [133]. Using data from HSQ, scores along the four dimensions of humor styles, which jointly denote the 'humor type' of individuals [133], were computed. We used clustering to group participants according to their humor type and then measured group differences in photo-sharing behaviors and reactions to interventions. Our analysis identified *humor type* as a significant predictor of photo-sharing behaviors, i.e., participants with different humor types exhibited significantly different likelihood of sharing memes. Moreover, humor type was significantly associated with past history of sharing embarrassing and privacy-violating photos of other people in real life. Finally, how the interventions influenced photo-sharing decisions depended on participants' humor type, i.e., participants with different humor type reacted differently when interventions were applied compared to the control condition. These findings shed light on the important role one's humor type plays in shaping their photo-sharing behaviors. They also establish humor type as an important factor to consider while designing interventions since advancing social connections are among the most important motivations for sharing photos online [151] and how people use humor to initiate or strengthen social relationships partly depend on their humor type [133].

7.2 Background

In this section, we define key terms and present background knowledge related to internet memes and people's humor style. We also justify why we focused on *individual humor style* and our

selection of the priming manipulations based on the prior literature.

7.2.1 Internet Memes and their Functions

Davison uses the following definition to describe internet memes: An Internet meme is a piece of culture, typically a joke, which gains influence through online transmission. [48]. Although memes can be text, image, or video based, we focus exclusively on ‘Image macros’ in our study, which was described by Rugnetta as “captioned images that typically consist of a picture and a witty message or a catchphrase. The structure of an image macro usually consists of a picture with text above and below the image in the macro” [175] and used in the literature [78]. Milner created a taxonomy of such memes and described their usage as multimodal artefacts to tell jokes, make observations, or advance arguments [140]. According to De la Rosa-Carrillo, internet users use memes to remix digital content to communicate jokes, emotions and opinions [115]. Grundlingh argues that memes function as a ‘speech act’ [78] and are predominantly used to communicate humorous and sarcastic contents but can also be used to communicate more serious contents and to ask questions [78].

7.2.2 Individual Humor Style

There have been a number of attempts to measure individual differences in appreciating, enjoying, and using humor (Martin *et al.* provide a review [133]). We use the classification system proposed by Martin *et al.* for two reasons – i) this widely-used measure had been validated by several other studies [59] and ii) the focus of this work was discovering individual differences in how people *use* humor to entertain themselves and/or other people to advance social relationships, which is particularly pertinent to the context of online photo-sharing. Martin and colleagues [133] identify four dimensions of humor style in their measure:

Affiliative humor: Individuals high in this dimension tend to use humor (e.g., jokes, spontaneous witty comments) to attract others’ attention, to entertain other people with the goal of

advancing social relationships, and to reduce interpersonal tensions. They are also likely to use light self-deprecating humor with a self-accepting tone to put others at ease, but may not use humor that are hostile to others.

Self-enhancing humor: Individuals high in this dimension usually possess a positive outlook towards life even in the face of difficulty. They use humor to entertain the self, sometimes as a strategy to cope with adverse situations. Thus, compared to Affiliative humor, the use of self-enhancing humor has a more personal than social focus.

Self-defeating humor: This dimension of humor style is socially-oriented, where individuals high on this dimension are likely to use self-disparaging humor (e.g., jokes about their weakness or funny things that make them look foolish) to gain approval from others and acceptance in a social circle. This dimension is also involved in the use of humor to hide underlying negative emotions.

Aggressive humor: This dimension of humor style relates to the use of sarcastic, ridiculous, and disparaging humor without regard for its potential impact on others. Individuals high on this dimension are also likely to make impulsive ‘jokes’ or say ‘funny’ things that may hurt others.

7.2.3 Relevance of ‘Humor style’ to Photo-sharing Behaviors

Scholars have extensively studied and established links between the humor styles and inter-personal skills to create and maintain social relationships [60, 178, 219], aggressive behaviors such as online trolling and cyberbullying [135, 162], other personality traits such as empathy and narcissism [80, 133, 210, 221], and demographic factors [92, 178]. These personality traits, in turn, were found to be associated with motivations to use social media platforms and sharing photos. For example, prior research suggests that humor style predicts social competence [219], empathy towards other people [80], and pro-social behaviors [60], while creating and maintaining social relationships

are among the primary motivating factors to share photos on social media [23, 40, 98, 99, 200]. Furthermore, focusing on memes, Preez and Lombard found that such photos partly shape the online persona one portrays on the social platforms [56]. Related to this result, Hunt and Langstedt documented that self-expression and self-presentation motivations were influenced by personality traits [98], which were in turn associated with the styles of humor [133]. Finally, trolling and cyberbullying behaviors, which are sometimes accomplished by posting memes, were found to be correlated with ‘maladaptive’ humor styles (*self-defeating humor* and *aggressive humor*) [135, 162].

7.2.4 Justifications of the Interventions

7.2.4.1 Engaging in Perspective Taking and Photo Sharing

Past research has shown that perspective taking – imagining that the self is in another’s position – can powerfully influence how one thinks about and behaves in a situation, often in service of prosocial goals. Based on this, we hypothesized that having participants take the perspective of the photo-subject may discourage them from sharing the memes, particularly if the memes portrayed the subjects in a negative light (i.e., negative valence). Perspective taking can create or increase self-awareness [187] and force people to view themselves through others’ eyes [187]. Much like looking at mirror can trigger self-awareness [38], thinking about how one would be viewed by others within a certain context (here, as the subject of a meme) may generate self-awareness. Higher level of self-awareness may result in more pro-social behaviors [20, 53, 70]. Further, taking others’ perspective has shown to reduce one’s prejudice and increase sympathy toward them [204, 209]. Thus, by imagining being the person in the shared photo, this may trigger both reputational concerns and sympathy for the person in the photo. Both of these may actually increase the sharing of photos that portray the subjects in a positive light but not the photos that portray the subjects negatively.

7.2.4.2 Adopting a Privacy Perspective and Photo Sharing

The second intervention explicitly asks the participants to consider the privacy of photo subjects. Our expectation that such priming would influence people's photo-sharing decisions is based on prior research that reported that people often show concern about others' privacy and refrain from sharing photos. For example, Jia and Xu observed collaborative behaviors of people on social media and found a tendency to collectively protect each other's privacy [218]. Recent work in the context of 'lifelogging' with wearable devices has shown that owners of those devices sometimes turn the devices off or delete photos afterward out of privacy concerns for the people in those photos [95]. These findings in the domain of photo sharing map closely onto broader work on the 'Sociology of Privacy,' as discussed by Anthony et al. [18]: across a variety of contexts, people often exhibit 'civil inattention' or what Goffman calls 'tactful inattention' [72] (i.e., purposely ignoring available information about others) and 'pretense awareness' (i.e., pretending not to know information about the other). For example, taxi drivers often pretend to not hear private conversations of their passengers (i.e., civil/tactful inattention), or one may ask questions of a new colleague, such as their dissertation topic, even though one has already closely read their application materials (i.e., pretense awareness). These behaviors highlight how people in society are willing to protect the privacy of others in public settings for the sake of propriety. With the Privacy Perspective intervention, we intend to capitalize on people's sense of 'propriety' by increasing awareness of privacy concerns at the moment of making a photo-sharing decision. Here again, any potential effects of a privacy-perspective intervention might also be qualified by the valence (i.e., the degree to which photos paint the subject in a 'positive' or 'negative' light) of the to-be-shared photos. If participants were made sensitive to the privacy of others, it is plausible that this would be especially true for more negative photos, as sharing negative information about others is a greater invasion of privacy than sharing positive information.

7.2.4.3 Time-delay for Better Decision Making

We implemented a time-delay of eight seconds as another intervention so that participants would have sufficient time to think before they make any decision regarding the sharing of the memes. Prior research has discovered that people may make poor decisions under time constraints [9, 144] and forcing them to spend more time to think before acting yields better outcomes. For example, Moser imposed a 25-hour delay before study participants could make online purchases, which significantly reduced impulse-buying [144]. Focusing on decision-making related to security, Volkamer *et al.* reported that when people were forced to wait for three seconds before they could click on links from phishing emails they were more likely to examine the link closely and less likely to click on it [212].

7.3 Method

We collected memes from the internet to use in our study. Prior to the main study, a separate online study was conducted to the valence ratings of those memes. The procedure to collect memes, their valence ratings, and the main study are described in the following sections.

7.3.1 Collecting Memes

One hundred and twenty publicly-available photo-memes were selected from popular social media sites (e.g., Facebook, Reddit, and Pinterest) and internet search engines. All of these photos included people and were accompanying text with 50 words or less, which provided context for photos. For example, one photo portrayed a woman and a man sitting together in a field surrounded by flowers, with text that read “Husband spends 2 years planting thousands of scented flowers for his blind wife to smell & get her out of depression.” Another photo included derogatory text directed toward a smiling subject with bad teeth, saying “9/10 dentists would recommend suicide.” From this initial set of photos, 98 were retained to be used in the subsequent study; the rest were discarded

as they either had very graphic content or overly offensive text, or included children among the photo-subjects. The remaining photos varied in terms of how the photo subjects were portrayed: some photo subjects were shown in ways that highlighted their accomplishments (e.g., completing a degree) or positive personal traits (e.g., performing an act of care). Other photo subjects were shown in ways that violated social norms (e.g., excessive alcohol intake) or highlighted negative personal traits (e.g., clumsiness).

7.3.2 Study I: Collecting Valence Ratings of the Memes

An online study was conducted to collect the valence ratings (i.e., the extent to which memes portrayed photo subjects as positive or negative) of each meme that was used in the second study.

7.3.2.1 Participants

Four hundred participants were enrolled from Amazon’s Mechanical Turk online recruitment system. Participants were eligible to participate in the study if they were 18 years or older, had been living in the United States for a minimum of five years, and used a laptop or desktop computer to complete the experiment. We followed the recommended procedures to minimize the chances that participants were not following our instructions [128]. This included restricting participation in the survey to workers who have at least 95% approval ratings and have completed at least 1,000 HITs.

One hundred and seventy-four (43.5%) and 221 (55.3%) participants identified themselves as female and male respectively. Participants were divided among four age groups: 150 (37.7%) were 18–29 years old, 210 (52.8%) were aged 30–49 years old, 25 were 50–64 years old, and 12 participants were 65 years or older. Sixty-seven percent (267) of the participants identified themselves as Caucasian, followed by Asian (53, 13.3%), Black or African American (31, 7.8%), American Indian or Alaska Native (19, 4.7%), and Hispanic or Latino (13, 3.3%). One hundred and sixty-nine participants (42.04%) had a Bachelor’s degree, 99 (24.63%) had some college education, 54 (13.43%) were high school graduates or had a GED, 41 (10.2%) had an Associate’s degree, and 38 (9.45%)

had a Master's degree. Participants had on average 3.2 ($SD = 1.58$) social media accounts. A majority of the participants (264, 66.3%) reported that they visit social media multiple times a day, and the frequency for sharing photos online had a mode response of 'multiple times a week' (96, 24.1%).

7.3.2.2 Procedure

After completing the informed consent form, participants viewed a sequence of 98 photos at the top center of their Amazon Mechanical Turk survey. One question was displayed below each photo, which asked participants "Does this portray the subject of the photo negatively or positively?" Participants provided ratings for each photo using a seven-point Likert scale (-3 = Extremely negative, -2 = Negative, -1 = Somewhat negative, 0 = Neither negative nor positive, 1 = Somewhat positive, 2 = Positive, 3 = Extremely positive). By rating the photo as positive or negative, participants were providing ratings of photo "valence," or the degree to which participants were portrayed in an aversive or bad way, versus an attractive or good way. Each participant was paid \$3 and the average time to complete the survey was approximately 36 minutes.

7.3.2.3 Results

Each photo was assigned a valence score by averaging ratings across participants. The means range from -1.74 to 2.45 for the 98 photos. The standard deviations ranged from 0.88 to 1.93 (Figure 7.1). Notably, the distributions of responses for each photo revealed that the means were not a function of a bimodal distribution of scores, but rather were a function of one or two consecutive Likert ratings constituting the most frequent response.

Photos were then ordered from most negative to most positive and divided into four quartiles with regard to how they portrayed subjects in the photos: very negative ($M = -1.15$, $SD = 0.34$, $N = 25$), negative ($M = -0.29$, $SD = 0.17$, $N = 24$), positive ($M = 0.38$, $SD = 0.23$, $N = 24$), or very positive ($M = 1.47$, $SD = 0.49$, $N = 25$). Note that, even though the valence categories we

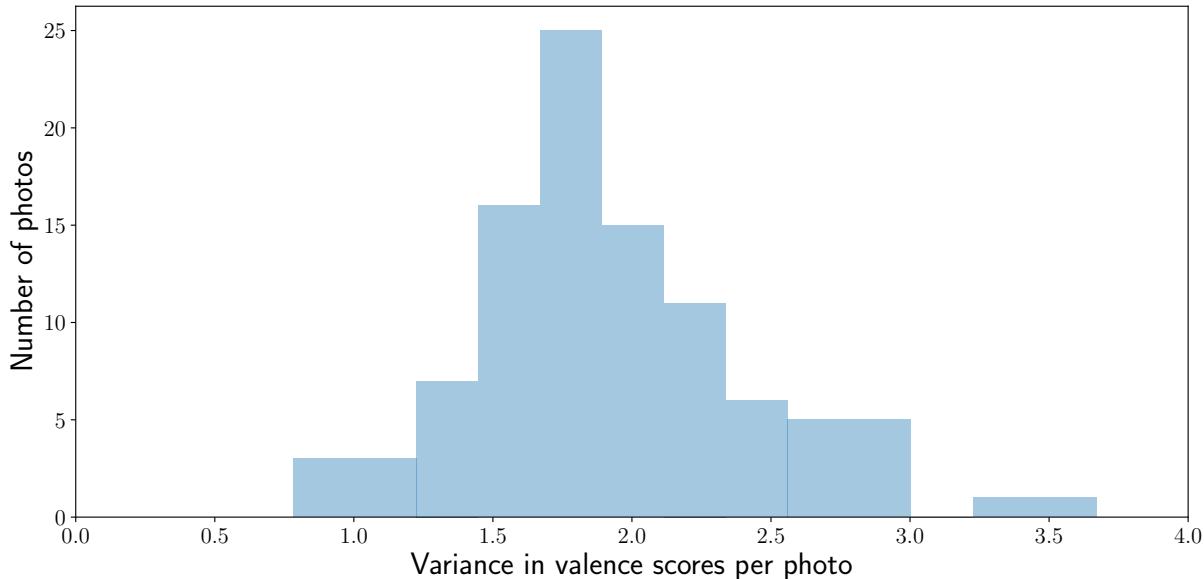


Figure 7.1: Histogram of variances for valence scores per photo.

used in the survey included the ‘Neutral’ valence (the *Neither negative nor positive*), after averaging the valence scores of each meme across participants, there was no meme with ‘Neutral’ category (i.e., zero mean score). Further, after grouping, all the memes in the *very negative* and *negative* groups had negative valence scores while those in the other two groups had positive valence scores.

Perceived valence ratings obtained from this study were used in the second study to differentiate photos into four valence categories to assess whether valence moderates likelihood of sharing responses.

7.3.3 Study II: Collecting Data on Photo-sharing Behaviors

7.3.3.1 Stimuli and experimental manipulation

Through an online survey, we collected data about participants’ likelihood to share memes under one of three experimental conditions. Data about participants’ social media usage and photo-sharing habits in real life were also collected. The same 98 memes collected by Amon *et al.* [16] were used in our experiment. In a pre-test study, the memes were rated by 400 participants according to how

Table 7.1: Questions Presented for Each Condition

Condition	Photo questions
Baseline	How likely are you to share this photo on social media?
Perspective-taking condition (PT)	<i>If this was a photo of you, how likely are you to share this photo on social media?</i>
Privacy-perspective condition (PP)	<i>Taking into account the privacy of the person in the photo, how likely are you to share this photo on social media?</i>

positively or negatively they portrayed the people appearing in them [16]. Average ratings across the participants for each meme indicates its ‘valence’ score (min=-1.74, max=2.45). These memes were then ordered according to the valence score and divided into four quartiles: very negative ($M=-1.15$, $SD=0.34$, $N=25$), negative ($M=-0.29$, $SD=0.17$, $N=24$), positive ($M=0.38$, $SD=0.23$, $N=24$), or very positive ($M=1.47$, $SD=0.49$, $N=25$). Participants in the present study viewed these memes in random order and were asked to indicate their preference to share these photos on social media. Table 7.1 shows the questions that were asked in the three experimental conditions. Two of them included priming manipulations by instructing the participants to *imagine themselves as the photo-subjects* (Perspective taking) and to *consider the privacy of the people in the photos* (Privacy perspective). These interventions were taken from [16], but in our experiment, we introduced a delay of eight seconds between displaying the meme (and corresponding question) and providing response options. The delay was added as an intervention test to see if Amon et al.’s paradoxical finding would be addressed by allowing for more time in decision making. A 7-point Likert scale was used to get their responses (-3 = Extremely unlikely to 3 = Extremely likely).

7.3.3.2 Questionnaires.

Four additional questionnaires were included in the study:

Social Media Usage Questionnaire. It assessed participants’ online social-media usage behavior including which social media platforms they had an account and how frequently they

visited those accounts and shared photos. Participants who shared photos online were further queried about how often they shared photos that were taken by themselves or people they knew (e.g., friends and family members) and photos taken by strangers or that were found on the internet (see Appendix C.1 for the complete questionnaire).

Social Media Privacy Questionnaire. This consists of 15 questions related to participants' online photo-sharing history and experiences related to privacy violations in real life. Eight questions asked about whether the participant had posted any photos of themselves and regretted afterwards (e.g., "Have you ever regretted posting a picture of yourself online?") or shared other people's (familiar or unknown) photos that may have violated their privacy (e.g., "Have you ever posted a picture of a stranger which may have violated his or her privacy?"). Four questions measured similar past behaviors of people known to the participants (e.g., "Has anyone you know posted a picture that may have violated someone's privacy?"). Finally, three questions asked whether the participants have been victims of privacy violations as a result of other people sharing their photos (e.g., "Has anyone ever shared a picture of you online that you felt violated your privacy?"). Answers were recorded on a three-point scale, either "no," "maybe," or "yes." Additionally, two attention check questions were included which instructed participants to provide a specific Likert-scale response (e.g., "Select the third option for this question.") or skip a question.

An additional Privacy Preference Question was administered, which asked participants, "Are you a private person who keeps to yourself or an open person who enjoys sharing with others (1 = very private, 7 = very open)?".

Humor Style Questionnaire. The *Humor Style Questionnaire* [133] was included to measure participants' humor styles. Each of the four dimensions of humor style was measured by eight questions, totaling to 32 questions. Participants responded using a 7-point Likert scale ('Totally disagree' = 1 to 'Totally agree' = 7).

7.3.3.3 Survey flow

First, the participants viewed the consent form containing study purpose, procedure, and payment information. After agreeing to participate, they completed the Social Media Usage Questionnaire. Next, they completed the experimental task, which required them to view all 98 photos one after another in a random order. Each photo was accompanied by a question asking about the likelihood of them sharing it on social media. In the perspective-taking and privacy-perspective conditions, a prime preceded the question. Unlike the study of Amon *et al.* [16], a delay of 8 seconds was introduced between the appearance of the photo (and accompanying question) and the appearance of the response options. We chose to delay the response by eight seconds based on an in-lab pilot study designed to determine the average amount of time necessary to decide on the likelihood of sharing the photo meme.

After answering questions for all photos, participants completed other questionnaires in this order: Social Media Privacy Questionnaire, Humor Style Questionnaire [133], Privacy Preference Question, and demographic questions (age, gender, racial background, and education level). They were included at the end of the survey to avoid priming the participants to think about privacy other than the interventions included in the experimental manipulations.

7.3.3.4 Participants

The surveys and questionnaires were implemented in Qualtrics¹ and participants were recruited through Amazon’s Mechanical Turk.² Workers who were at least 18 years old and had been living in the United States for a minimum of five years were eligible to participate in the study. The study was further restricted to workers who had completed at least 1,000 HITs and had at least 95% approval ratings to ensure data quality [128]. To ensure proper viewing of the photos, participants were required to use a laptop or desktop computer to answer the survey questions. Of the 556

¹<https://www.qualtrics.com>

²<https://www.mturk.com>

respondents, 437 responded correctly to both attention checks and were retained for the final sample; responses from the remaining participants were discarded. Eighty-two (18.7%) participants were between the ages of 18–29 years, 278 (64%) were between 30–49 years, 70 (16.5%) were between 50–64 years, and seven (1.6%) were older than 65 years. One hundred and ninety-two participants (43.9%) identified as female and 244 (55.8%) identified as male. Three hundred and fifty-eight participants (76.99%) identified themselves as Caucasian, 41 (8.8%) as Black or African American, 30 (6.5%) as Asian, 28 (6%) as Hispanic or Latino, seven (1.5%) as American Indian, and 1 (0.22%) as biracial or multiracial or “other.”. Participants ranged in education from having some high school education (11.%) to having doctoral (0.23%) or professional degrees (1.1%). The mode for education level was a bachelor’s degree (38.9%), followed by having completed some college (26.7%), followed by Associate’s degree (13.8%), and then high school or GED (11%). Most of the participants (97%) reported having at least one social media account and the average number of accounts was 4.20. . The majority of participants visited ‘multiple times per week’ ($n = 341$, 72.3%). On average, the participants share photos on more than one social media and almost one-third of them ($n = 129$, 29%) share photos ‘multiple times per week’. A majority of participants (54%) share photos with familiar people while the rest share photos publicly. Participants were randomly assigned to one of the three experimental conditions: 150 participants were in the *Baseline* condition, 141 participants were in the *Perspective-taking* condition, and 144 participants were in the *Privacy-perspective* condition. The median completion time for the survey was 37 minutes while 75% of the participants completed it within 49 minutes. All participants who completed the survey were paid \$5 regardless of whether their data was used for analysis or not. The study protocol was reviewed and approved by our institution’s ethics review board for the protection of human subjects.

Humor styles of the participants. In our sample of data, participants had similar means and standard deviations along the four dimensions of humor style as the original study by Martin *et al.*

[133]: Affiliative ($M=43.9$, $SD= 8.7$), Self-enhancing ($M=40.3$, $SD=9.2$), Self-defeating ($M=28.4$, $SD=9.7$), and Aggressive ($M=25.5$, $SD=9.1$). There was no significant difference in scores along any of the dimensions among the three experimental conditions (all $p > 0.05$).

7.3.4 Methods of Data Analysis

7.3.4.1 Validating HSQ and clustering participants based on humor styles

First, we validated the Humor Style Questionnaire using confirmatory factor analysis. Our experimental data supported the four-factor structure representing four dimensions of individual humor styles. While each of these dimensions indicates a single aspect of how one expresses humor, all four dimensions have to be considered simultaneously to get the full picture of one's humor type. Recently, researchers have been critical of the practice of studying how each of these dimensions independently correlate with other personality traits and behaviors [59, 119]. They advocated for grouping people by simultaneously considering all four dimensions of humor and then looking into group differences [59, 119]. This approach has been adopted by more contemporary studies [59, 66, 119]. In particular, Evans and Steptoe-Warren reported that humor clusters are better predictors of individual differences in communication, stress level, and creativity, than the humor styles [59]. We followed this recommendation and used K-Means [158] to cluster the participants based on their scores along the four dimensions of humor style. The number of clusters (K) were determined experimentally by examining the error in the model for different values of K . For each configuration, the sum-of-squared distances among the data samples and their closest cluster center represents the coherence of the cluster and is plotted in Fig. 7.2. Based on the ‘elbow-method’ [75], we identified a three cluster configuration as the best configuration. This is what was also reported by several prior studies [59, 119], providing further evidence in support of this cluster structure.

In our case, there were 176, 113, and 148, people in the three clusters, respectively. To interpret these clusters based on the four dimensions of humor, the z-scores of the cluster means along those

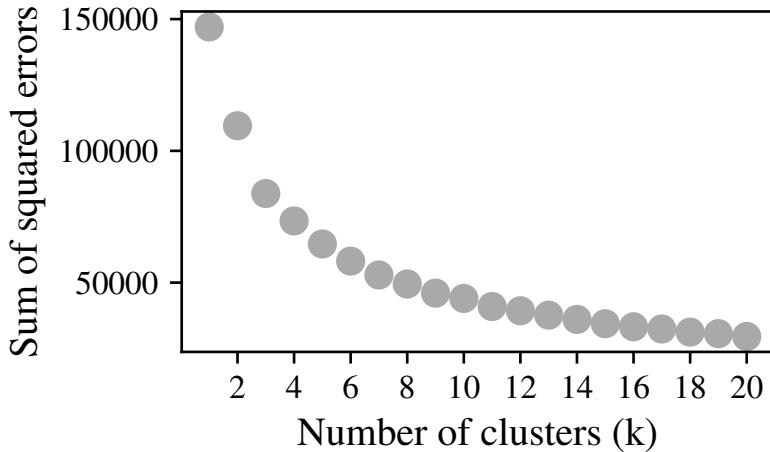


Figure 7.2: Sum of squared distances of data samples to their closest cluster center for different number of clusters. The number of clusters was set to three based on the elbow method [75].

dimensions are shown in Table 7.2. Cluster 1 has an above-average amount of all four humor styles, while Cluster 2 has below average scores in all humor styles. The third cluster has above-average scores for the ‘Affiliative’ and ‘Self-enhancing’ sub-scales but below-average scores for the ‘Self-defeating’ and ‘Aggressive’ sub-scales. Notably, the properties of these three clusters are strikingly similar to those discovered in prior works [59, 119]. We therefore followed Leist and ’Müller [119] and labeled the three clusters: *humor endorsers* (female=27%), *humor deniers* (female=43%), and *self-enhancers* (female=64%).

Table 7.2: Z-scores of the cluster means along the four dimensions of humor.

	Cluster1	Cluster2	Cluster3
Affiliative	0.31	-1.16	0.52
Self-enhancing	0.25	-1.12	0.55
Self-defeating	0.75	-0.40	-0.59
Aggressive	0.74	-0.17	-0.75

7.3.4.2 Statistical analyses

Different statistical models were utilized to answer different research questions. To analyze data about meme-sharing likelihood under different experimental conditions, a mixed linear model was

built where the likelihood to share a meme was the dependent variable and *experimental condition*, *photo valence*, *humor types*, and interaction terms involving them were used as the predictors. We controlled for *gender* and *age*. All pairwise comparisons (with appropriate method for p-value correction) were performed using the estimated means obtained from this model.

To examine the extent to which meme-sharing behaviors under the controlled experimental setup is associated with the *real life* photo-sharing behaviors, we performed correlational analyses using average meme-sharing likelihood of a person and their responses to the questions asking about incidents of sharing privacy-sensitive photos (*Social Media Privacy Questionnaire*). Additionally, we tested whether real-life photo-sharing habits vary as a function of humor type by building logistic regression models using responses to *Social Media Privacy Questionnaire* and conducting Likelihood Ratio tests. In all cases, responses to the questions were binary coded ('Yes' = 1, 'No' = 0) after removing the uncertain ('Maybe') responses.

Finally, several linear regression models were built to analyze the effect of humor type in social media usage and *generic* photo-sharing behaviors (i.e., not restricted to privacy-sensitive photos). More specifically, separate models were built with *number of social media accounts*, *frequency of sharing own photos*, and *frequency of sharing other people's photos* as the outcome variables and 'humor type' as the predictor. In each case, we controlled for *age* and *gender*.

7.4 Findings

7.4.1 Relation Between Humor Type and Photo-sharing Behaviors

Table 7.3 shows the results from the omnibus test involving the mixed effect model: all of the predictors of primary interests – *humor cluster*, *experimental condition*, and *photo valence* – and some interaction terms involving them had significant effects on photo-sharing likelihood. In the following sections, we will state key takeaways from the findings, backing up with statistical evidence from Table 7.3 and results from additional pairwise comparisons.

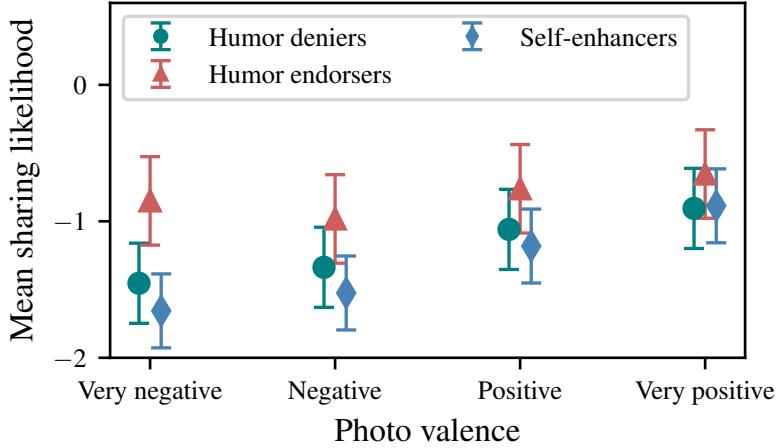


Figure 7.3: Mean (with 95% CI) sharing likelihood by humor type and photo-valence in the *baseline* condition.

Finding 1. Without any behavioral manipulations, *humor endorsers* are significantly *more likely* to share *very negative* memes than *humor deniers* and *self-enhancers*. As shown in Table 7.3, *humor cluster* significantly predicts meme-sharing likelihood ($F(2, 436) = 8.68, p < 0.001$), but this effect is superseded by a higher order interaction effect involving *condition* and *valence* ($F(12, 42292) = 2.97, p < 0.001$). This suggests that people in different humor clusters differ in meme-sharing likelihood depending on the valence of the meme and the experimental condition. We conducted pairwise comparisons using data in the *Baseline* condition to quantify how people in different humor clusters share photos in different valence groups *without any behavioral manipulations*. We found that, for *very negative* photos, *humor endorsers* demonstrated significantly higher sharing likelihood ($M = -0.85, SE = 0.165$) than both *self-enhancers* ($M = -1.66, SE = 0.14$) and *humor deniers* ($M = -1.45, SE = 0.150$), $p = 0.03$ and $p < 0.0001$, respectively (also see Fig. 7.3). All other comparisons were non-significant (all $p > 0.05$). In summary, people who frequently make use of humor either to enhance themselves or entertain others are also more likely to share memes that negatively portray other people and thus may violating the photo-subjects' privacy.

Finding 2a. Participants' meme-sharing likelihood during the experiment is significantly correlated with their past history of photo-sharing behaviors on social media.

Table 7.4 shows the (Pearson's product-moment) correlation coefficients between the average sharing likelihood (across all the memes) of a participant and their responses to the questions that asked whether they have shared embarrassing or privacy-violating photos of themselves or others. All coefficients are statistically significant, suggesting that findings about their meme-sharing behaviors in the experimental settings may generalize to their *real life photo-sharing behaviors*.

Finding 2b. Humor type was not a significant predictor of whether participants had shared *privacy-sensitive* photos of themselves or others in real life. Likelihood Ratio tests involving logistic regression models revealed that only in one case ("Have you ever posted a picture online of someone else you know, which may have violated his or her privacy?") participants' humor type affected their behaviors ($\chi^2(1) = 7.8, p = 0.02$). But there was no significant difference in the odds of participants responding "Yes" across the humor types.

Finding 2c. People differed in terms of social media usage and the sharing of *generic* photos depending on their humor type. *self-enhancers* and *humor endorsers* were more engaged in social media usage and photo-sharing activities. Humor type was a significant predictor for how many social media accounts participants had ($F(2) = 3.53, p < 0.05$), how frequently participants visited those accounts ($F(2) = 4.52, p < 0.05$), how frequently they shared photos of themselves ($F(2) = 7.81, p < 0.001$), and how frequently they shared photos of other people ($F(2) = 4.59, p < 0.05$). Posthoc tests with Dunnet's method for p-value adjustment revealed that *self-enhancers* ($M = 4, SE = 0.14$) had more social media accounts than *humor deniers* ($M = 3.6, SE = 0.17$), $p < 0.05$. Self-enhancers also visited their accounts more frequently ($M = 6.6, SE = 0.12$ than *humor deniers* ($M = 6.1, SE = 0.15$), $p < 0.05$. Both *humor endorsers* ($M = 3.1, SE = 0.20$) and *self-enhancers* ($M = 3.5, SE = 0.18$) shared photos of themselves more frequently than *humor deniers* ($M = 2.5, SE = 0.23$), $p < 0.05$ and $p < .0001$, respectively.

But only *humor endorsers* ($M = .4$, $SE = 0.24$) shared photos of other people more than *humor deniers* ($M = 2.6$, $SE = 0.25$), $p < 0.05$. All other comparisons were non-significant at the 95% significance level.

7.4.2 Reactions to the Interventions

In this section we present results related to the effects of behavioral interventions on the participants and whether they differed depending on their membership to different humor clusters.

Finding 3a. Both behavioral interventions (paradoxically) resulted in *higher* likelihood of meme-sharing. In other words, people were likely to share *more* when they imagined themselves as the photo subjects (PT condition) and when they were reminded about others' right to privacy (PP condition). We found a significant main effect of *condition* ($F(2, 436) = 5.06$, $p = 0.00674$), indicating that the interventions influenced meme-sharing behavior (Table 7.3). Pairwise comparisons among the conditions revealed that participants in the *Privacy perspective* ($M = -0.80$, $SE = 0.090$) and *Perspective taking* ($M = -0.76$, $SE = 0.096$) conditions share significantly more than participants in the *Baseline* condition ($M = -1.10$, $SE = 0.089$) ($p < 0.05$ in all cases),³ confirming the paradoxical effect of the interventions observed by Amon *et al.* [16].

Finding 3b. Only *humor deniers* increase sharing likelihood when reminded about others' privacy (PP condition). People in other humor clusters do not exhibit this paradoxical behavior. Following the significant three-way interaction involving humor, condition, and valence (Table 7.3), we conducted post-hoc pairwise comparisons to further investigate how the *Perspective taking* intervention affect members of different humor clusters for photos across the valence groups.⁴ Analyses revealed that only *humor deniers*, who scored less than average along all four dimensions of humor (i.e., they use any type of humor *infrequently*), increases

³after adjusting p-value using Dunnett's method

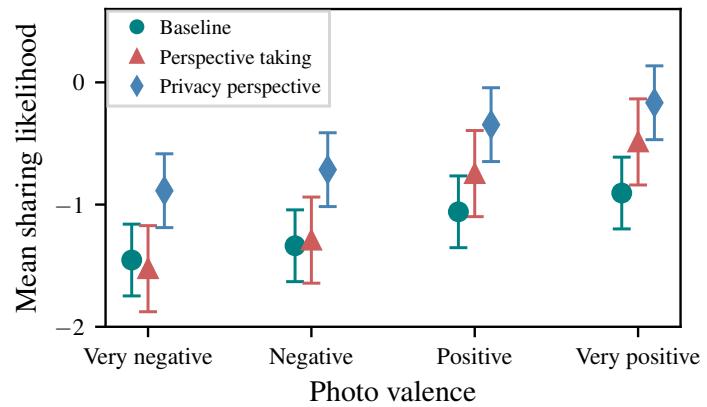
⁴Dunnett's method was used for p-value correction.

sharing likelihood in response to the intervention, and this is true regardless of the photo-valence (Fig. 7.4a). In other words, *humor deniers increased* sharing memes belonging to all valence groups ($M_{very_neg} = -0.89$, $SE_{very_neg} = 0.15$, $M_{neg} = -0.71$, $SE_{neg} = 0.15$, $M_{pos} = -0.35$, $SE_{pos} = 0.15$, $M_{very_pos} = -0.17$, $SE_{very_pos} = 0.154$) in the *Privacy perspective* condition compared to the *Baseline* condition ($M_{very_neg} = 1.45$, $SE_{very_neg} = 0.15$, $M_{neg} = -1.34$, $SE_{neg} = 0.15$, $M_{pos} = -1.06$, $SE_{pos} = 0.15$, $M_{very_pos} = -0.91$, $SE_{very_pos} = 0.15$), all $p < 0.05$ (see Fig. 7.4a). There was no significant effect of this intervention on the *humor endorsers* and *self-enhancers*.

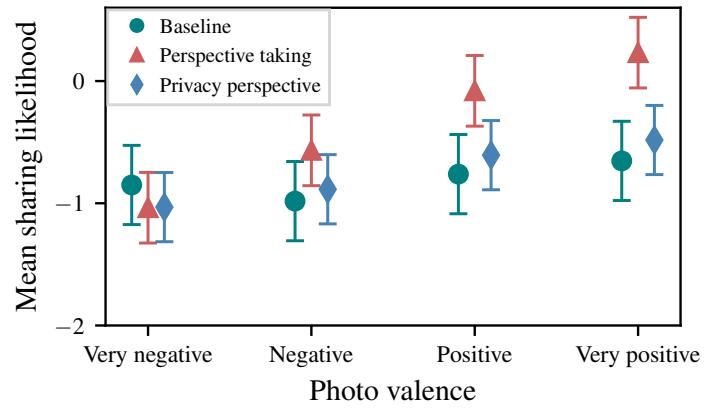
Finding 3c. Both *humor endorsers* and *self-enhancers* increased sharing likelihood when they imagined themselves as the photo-subjects (*PT* condition), but only when the photos portrayed the subjects *positively* (i.e., positive valence). Participants in the *humor endorsers* cluster demonstrated significantly *higher* sharing likelihood for *positive* ($M = -0.08$, $SE = 0.15$) and *very positive* ($M = 0.23$, $SE = 0.15$) photos in the *Perspective taking* condition compared to the *Baseline* condition ($M_{pos} = -0.76$, $SE_{pos} = 0.17$, $M_{very_pos} = -0.65$, $SE_{very_pos} = 0.17$), $p < 0.0001$ in both cases (see Fig. 7.4b). On the other hand, *self-enhancers* increased sharing likelihood significantly in *Perspective taking* condition ($M = 0.77$, $SE = 0.14$) only for *very positive* photos compared to the *Baseline* condition ($M = -0.89$, $SE = 0.14$), $p < 0.0001$ as shown in Fig. 7.4c.

7.4.3 Effect of Gender

Finding 4. Females demonstrated *higher* sharing likelihood than males for *positive* and *very positive* photos regardless of humor group and experimental condition. Gender has significant effect on photo-sharing likelihood ($F(1, 436) = 6.91$, $p = 0.009$), but this effect is moderated by photo-valence ($F(3, 42292) = 53.5$, $p < 0.0001$), as shown in Table 7.3. Pairwise comparisons revealed that female identifying participants were significantly more likely to share *positive* and *very positive* ($M_{pos} = -0.55$, $SE_{pos} = 0.13$, $M_{very_pos} = -0.15$, $SE_{very_pos} = 0.13$)



(a) Humor deniers



(b) Humor endorsers

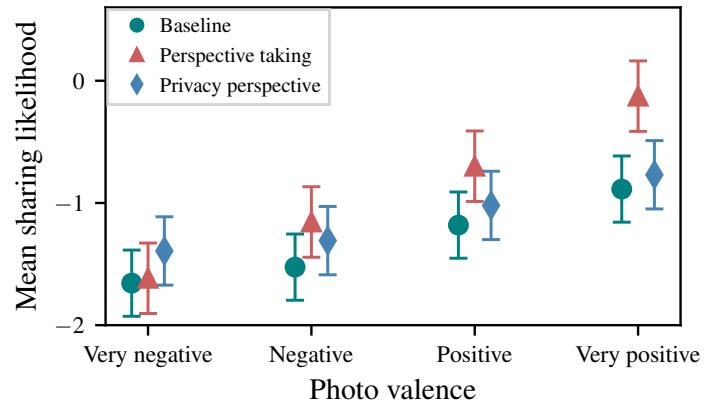


Figure 7.4: Means (and 95% CI) of sharing likelihood

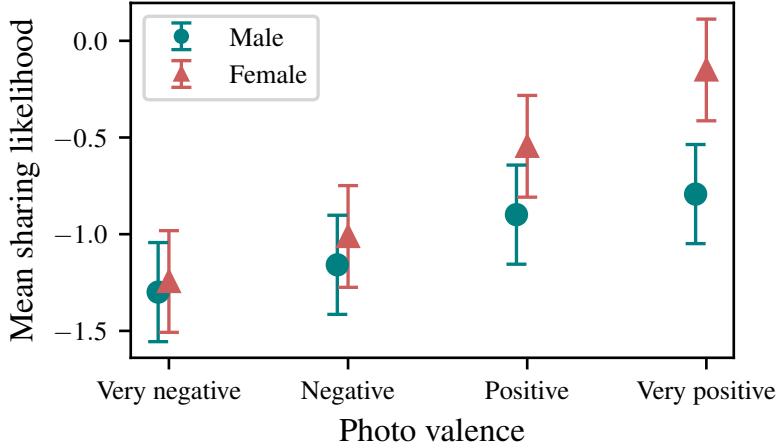


Figure 7.5: Mean likelihood (with 95% CI) to share photos by Females and Males across valence levels.

photos compared to male identifying participants ($M_{pos} = -0.90$, $SE_{pos} = 0.13$, $M_{very_pos} = -0.79$, $SE_{very_pos} = 0.13$), $p < 0.0001$ in both cases (also see Fig. 7.5). All other comparisons were non-significant ($p > 0.05$).

7.5 Discussion

One of the primary goals of this research was to assess the effects of individual differences in humor style on making decisions to share others' privacy-sensitive photos online. Our findings provide considerable evidence that humor style is an important predictor of photo-sharing behaviors in real life. Another primary focus of this work was to investigate whether people with different humor types react differently to behavioral interventions that were designed to discourage the sharing of privacy-sensitive photos. We reproduced the paradoxical result reported by Amon *et al.* [16] – the privacy interventions resulted in a higher sharing likelihood – and we could also pinpoint the sub-population that is more likely to exhibit this unexpected behavior. Finally, we investigated how humor type interacted with gender, but the interaction effect was not significant and our result on the effect of gender was not aligned with what was reported by Amon et al. [16]. We interpret

these findings below.

7.5.1 *Humor Endorsers* are More Likely to Share Memes with *Very Negative Valence*.

Humor endorsers have above average scores along all dimensions of humor styles and are characterized by frequent use of humorous content to entertain themselves or other people. But why did they differ from other humor groups for only the *very-negative* memes? Referring back to Fig. 7.3, it can be seen that, the difference was created because *self-enhancers* and *humor deniers* displayed a lower likelihood of sharing *very negative* memes, and not because *humor endorsers* shared *very negative* memes at a higher rate than other memes. This was expected since *humor endorsers* frequently use both positive and negative humor. In fact, as shown in Table 7.2, *humor endorsers* are further from the mean along the *self-defeating* and *aggressive* dimensions compared to the other two dimensions of humor style. These two dimensions (i.e., *self-defeating* and *aggressive*) are related to the usage of *negative* or *disparaging* humor [133]. Thus, *humor endorsers* concentrated on the humorous aspects of the memes even if the photo-subjects are portrayed negatively by those memes, and expressed their intention of sharing them at the same level as *positive* memes. On the other hand, *humor deniers* and *self-enhancers* are less likely to use negative humor (Table 7.2) and thus they lowered their sharing likelihood for *very-negative* memes.

7.5.2 Reactions to the *Perspective Taking* Intervention.

When *humor endorsers* and *self-enhancers* took the perspectives of the photo-subjects, they increased the sharing likelihood, but only for photos that portrayed the subjects in a positive light (Fig. 7.4b and Fig. 7.4c). Choosing the memes to share on social platforms is informed by the type of online persona one tries to create [56], thus it is not surprising that participants shared more when they imagined themselves as the photo-subjects and who were portrayed positively. But

this effect was observed for *humor endorsers* and *self-enhancers* and not for *humor deniers*. One possible explanation is that *humor endorsers* and *self-enhancers* have above average scores on the affiliative dimension of humor, which is correlated with high level of narcissism or overly positive self-view [132, 210], which in turn is associated with presenting the self in a positive light [37, 153]. Further, narcissistic people are more likely to share selfies (i.e., photos containing themselves) on social media [213] to gain others' attention [23]; thus imagining themselves in the memes resulted in a higher sharing-likelihood. It is worth noting that the *Perspective Taking* intervention was originally intended to *lower* the sharing of memes by increasing empathy towards the photo-subjects, but this surprising effect of *increasing* the sharing likelihood was also observed in that study [16] (but only for the *very positive* memes). The authors explained this phenomenon as a form of pro-social behavior by the participants, inspired by self-reflection and putting themselves in another's place, where they helped the photo-subjects to create a positive online persona by sharing their photos that were portrayed positively. Looking at this phenomenon through the lens of humor style, self presentation, and advancing social relationships provide an alternative explanation. Participants who are interested in positive self-presentation and enhancing social relationships increased sharing of photos that they imagined presented themselves in a 'good' way to their social connections, rather than treating it as a pro-social act (e.g., helping others to build positive persona) or an anti-social act (e.g., violating others' privacy by sharing their photos without their consent). This explains why *humor deniers*, who are less interested in advancing social connections, did not increase sharing of memes in the PT condition.

7.5.3 Reactions to the *Privacy Perspective* Intervention.

Participants in the *humor deniers* group *increased* photo-sharing likelihood when they were reminded about the photo-subjects' privacy (PP condition) compared to the control group, but *self-enhancers* and *humor endorsers* did not demonstrate this pattern. This paradoxical effect was

also reported by Amon *et al.* and the authors provided some hypotheses as to why that happened, including i) feeling more in control and thus more comfortable to share others' personal information, ii) explicitly rejecting the values of the intervention and, iii) reactance or the tendency for apparently unnecessary rules to elicit the opposite effect as intended. These hypotheses are not supported by our findings since we saw the paradoxical effects only for one group of the participants. Why only *humor deniers* behaved paradoxically? One plausible explanation may be narrowing decision criteria through priming. There are many reasons to (not) share memes online including funniness, appropriateness, relating to the self, and eliciting social interactions [16, 139], e.g., likes and comments. Thus, one might consider multiple reasons before deciding to share a meme, or not share when one or more of the conditions were not satisfied (e.g., a meme may be funny but not appropriate [16]). Since *humor deniers* are neither very appreciative of humorous content nor interested in using humor to advance social relationships – not satisfying many of the reasons to share memes – they are less likely to share memes in the control condition. But when they were warned about possible privacy implications of the sharing act, their decision to share the meme was based on only whether it will violate the photo-subjects privacy.

As reported by Amon et al., participants did not consider sharing the memes will violate the subjects' privacy for many reasons, including *the memes were already public* and *the subjects would not take the photos if they did not want them to be shared* [16]. Thus, deciding based on only this criterion, it seems reasonable that the sharing would increase. In other words, the priming narrowed the participants' attention and they did not explore all the reasons to (not) share the meme. Past psychological research supports the above hypothesis. For example, Friedman *et al.* showed that a narrow (broad) scope of perceptual attention results in an analogously narrow (broad) focus of conceptual attention [65], which in turn restricts (expands) the diversity of thoughts. A great deal of research has shown that deliberation can result in poorer judgment and decision-making compared to using intuition (Dijkstra et al. provide a review [54]). In our case, with the priming,

the *humor deniers* were forced to think about privacy, hindering their spontaneous reaction about whether to share a meme (which is most often *not sharing*).

Why self-enhancers and humor endorsers did not exhibit this paradoxical behavior after the same intervention? One possible reason is that both *self-enhancers* and *humor endorsers* are more appreciative of the humorous and social aspects of sharing photos and thus the priming had a smaller impact on narrowing their thoughts. Alternatively, both *self-enhancers* and *humor endorsers* score high along the affiliative and self-enhancing dimensions of humor, which are correlated with social competence [219] and pro-social behaviors [60]. Thus, *self-enhancers* and *humor endorsers* are more likely to consider the negative impact of violating someone's privacy (pro-social behavior) and how the memes will be received by their connections on the online platforms where often the photo-subjects are portrayed in embarrassing manners (social competence).

Why the perspective-taking intervention did not similarly affect the humor deniers (i.e., narrowing their focus)? One possible reason is that PT encourages one to relate to the photo-subject or the story depicted by the meme, and present oneself in an entertaining way to their social connections. But *humor deniers* are less likely to exhibit empathetic behaviors [80] and by definition, they are not interested in using (self-referential) humor to entertain others.

7.5.4 Effects of Gender.

Our results suggest that women as opposed to men tend to share more when the memes portray the subjects as *very positive*. This result deviates from what Amon et al. [16] reported: women demonstrated *lower* likelihood of sharing *negative* memes than men but no difference was found for other valence groups. Our result is consistent with prior research demonstrating that women engage more with online social media [94, 94, 138, 143, 143] and post photos more frequently than males. It is also in line with the heightened concerns about self-privacy [94, 182, 202] and risk-averse behaviors [35, 39] of women: memes that portray the photo-subjects in a positive light and

sometimes offer constructive messages of social interest may enhance their online reputation rather than harming their privacy and social impression. We did not find any significant interaction effect involving humor and gender, suggesting that people in the same humor group exhibit similar photo-sharing patterns regardless of their gender.

7.5.5 Effect of Time Delay.

During the experiment, we required participants to view the memes for eight seconds before they could respond to indicate sharing likelihood. The questions in each experimental condition were the same as [16], but in that study, participants could provide their responses immediately after viewing the meme. To quantify the effect of this time delay, we obtained the data reported in [16] and compared with our own. In particular, we compared the mean sharing likelihood of participants in the ‘Baseline’ condition of both experiments. The difference was not significant, indicating that the time delay did not alter participants’ meme-sharing preference.

7.5.6 Limitations

There were some limitations to this study that we discuss here. First, we collected data from workers on Amazon’s Mechanical Turk (MTurk) platform, who are known to be more privacy-concerned than the general US population [105]. But a study has shown that, in the context of conducting surveys concerning security and privacy, MTurk participants resemble the US population fairly well and better than other web panels [171]. In this experiment, participants’ scores along the four dimensions of humor style were comparable to the results reported by Martin *et al.* [133], who administered this questionnaire on a sample of undergraduate students in Canada. The clusters (denoting humor types) identified from this data were similar to prior studies conducted on participants from Germany [119] and United Kingdom [59] who were recruited through multiple methods including in-person, e-mail, and social media, providing further assurance regarding the generalizability of our findings. To reduce noise and maintain data-quality, we removed responses

from participants who provided wrong answers to any of the two attention check questions.

Second, in our study, we collected data about sharing image macros or memes. Preference to share such photos may differ from sharing photos without captions or any other types of alteration. Further, participants viewed and made sharing decisions for 98 photos in a row, which is not the usual when people view and share memes. But many participants, when asked to comment about the study, mentioned that they enjoyed the memes they saw and we did not find any indication of fatigue or boredom. Further, Amon *et al.* found no order effect in their data (that was collected in a similar experimental setting using the same set of memes), i.e., participants were engaged to the study from beginning till the end and consistently answered all questions. Regarding photo-sharing behaviors in real life, we collected self-reported, memory-based data which is prone to biased responses [214] (most relevantly *confirmation bias* and *consistency bias*) and may not be reliable. However, data about participants' past history of social media usage and photos-sharing frequency, and meme-sharing preferences during the experiment were in agreement with each other and were consistent with expected behaviors according to their humor types.

7.6 Conclusions

We investigated how individual humor style, which has been linked to many personal characteristics relevant to social media usage (e.g., social competence), affects photo-sharing behaviors on online platforms. We found that, humor style not only predicted participants' likelihood to share memes during our study but also was associated with their usage of social media in real life and past history of sharing privacy-sensitive photos of other people. In particular, participants who frequently use aggressive and self-disparaging humor were more likely to share memes and have shared photos in the past that may have violated others' privacy. Moreover, participants who infrequently use humor demonstrated the paradoxical behavior of sharing memes at a higher rate after they were primed to consider the photo-subjects' privacy. We discussed possible reasons behind this phenomenon,

which may guide future research in this direction. Our findings will help to develop effective and personalized behavioral interventions based on the humor style of the recipients to discourage them from sharing photos that may threaten others' privacy.

	Sum Sq	Mean Sq	Num DoF	Den DoF	F statistic	η_p^2
gender	20.21	20.21	1.00	436.00	6.91**	< 0.01
age	0.54	0.54	1.00	436.00	0.19	< 0.01
others' photo sharing frequency	98.45	98.45	1.00	436.00	33.64***	< 0.01
privacy perception	82.79	82.79	1.00	436.00	28.29***	< 0.01
condition	29.61	14.81	2.00	436.00	5.06**	< 0.01
valence-group	3776.53	1258.84	3.00	42292.00	430.06***	0.03
humor cluster	50.82	25.41	2.00	436.00	8.68***	< 0.01
gender : condition	0.60	0.30	2.00	436.00	0.10	< 0.01
gender : valence-group	469.84	156.61	3.00	42292.00	53.50***	< 0.01
gender : humor-cluster	10.07	5.04	2.00	436.00	1.72	< 0.01
condition : humor-cluster	24.27	6.07	4.00	436.00	2.07	< 0.01
condition : valence-group	721.83	120.31	6.00	42292.00	41.10***	< 0.01
valence-group :						
humor-cluster	81.42	13.57	6.00	42292.00	4.64**	< 0.01
condition : valence-group :						
humor-cluster	104.36	8.70	12.00	42292.00	2.97**	< 0.01

Table 7.3: Type II ANOVA Table (with Satterthwaite's method). (* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$).
The effect size η_p^2 (partial η^2) can be interpreted as small if $\eta_p^2 = 0.01$, medium if $\eta_p^2 = 0.06$, and large if $\eta_p^2 = 0.14$ [117].

#	Question	<i>r</i>
1	Have you ever regretted posting a picture of yourself online?	0.11*
2	Have you ever accidentally posted a picture of yourself online that you did not want to share?	0.19
3	Have you ever shared an embarrassing picture online of someone else you know?	0.16**
4	Have you ever regretted posting a picture online of someone else you know?	0.14**
5	Have you ever posted a picture online of someone else you know, which may have violated his or her privacy?	0.11*
6	Have you ever shared an embarrassing picture online of a stranger (someone that you do not personally know)?	0.23***
7	Have you ever regretted posting a picture online of a stranger (i.e., someone you do not personally know)?	0.14**
8	Have you ever posted a picture of a stranger (i.e., someone you do not personally know), which may have violated his or her privacy?	0.10*

Table 7.4: Correlation between the average meme-sharing likelihood of a participant and their past activities of sharing embarrassing or privacy-violating photos of themselves or others on social media (* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$).

CHAPTER 8

Discussions, Limitations, and Future Work

This dissertation makes significant contributions in reducing people’s privacy risks when their photos are shared on social media by others. We focused on two scenarios: i) protecting bystanders’ privacy, and, ii) protecting people’s privacy who became meme subjects. Although in both cases people were subjected to privacy violations when their personal information (e.g., identity, location, and activity) was revealed without their consent (and sometimes awareness), we approached them differently. To protect bystanders’ privacy, we devised a machine learning based system that can automatically distinguish them from subjects. We also studied how effective and usable image filters are in protecting bystanders’ privacy. These filters can be automatically applied over bystanders in images once they are identified by our system. To protect meme subjects’ privacy, we designed behavioral interventions to appeal to photo-sharers’ sense of ‘proprietary’ and discourage the sharing of memes. These two approaches can be combined to form a holistic ‘socio-technical’ approach to alleviate the privacy risks of both bystanders and meme subjects.

For example, behavioral interventions can be employed to inspire the usage of technical solutions to protect bystanders’ privacy. Although the bystander detector and image obfuscations can be seamlessly integrated into social media platforms, and as a result, privacy-enhanced versions of photos can be shared on these platforms without disrupting users’ usual workflows, the filtered images may not be always satisfactory to the owners (as shown in chapter 5). Thus, behavioral interventions may be employed to encourage the photo sharers to protect bystanders’ privacy at the expense of a slight loss in utility. We studied priming manipulations in a meme-sharing context where memes are primarily used to express humorous content (section 7.2.1). In the study, we investigated how people with different propensity to use humor for self-entertainment or to advance social relationships differ in meme-sharing behaviors and react to privacy nudges. Thus, the findings

may directly apply to the situations when people share photos they own (which contain bystanders) for entertainment purposes, although the ownership relationship may moderate the outcomes. Many participants of our study justified their decisions to share memes by noting that *the memes were already public* and *other people would share them even if they do not*; these justifications are not applicable in the case of sharing own photos. Thus, we expect that behavioral interventions would be more effective in encouraging privacy-respecting behaviors in the latter case even though they elicited the opposite behaviors than expected in our experiment. Indeed, past research has shown that photo owners were careful not to share photos containing people they did not know [95]. Moreover, the goal here is to encourage people to obfuscate bystanders before sharing their photos, as opposed to preventing the sharing of photos altogether, as in the case of memes. This may make the perceived expense of adhering to the intervention ‘acceptable.’ We also propose several ‘visual primings’ in the future work section, which may be more easily implemented in photo-sharing applications, and are expected to be more effective than text-based manipulations.

On the other hand, the bystander classifier may be utilized to reduce privacy risks in the meme-sharing context. The focus here would be to detect photo subjects (as opposed to bystanders), who are usually maligned with added captions, and then obfuscate their identity or other sensitive attributes by, e.g., using image filters.

Thus, combining the proposed social solutions to motivate social media users in adopting privacy-respecting and prosocial behaviors with the technical means to share information on social media in a privacy-preserving manner, we can go one step further toward a holistic solution to protect people’s privacy, who are victimized when other people share their photos on social media.

8.1 Limitations and Future Work

8.1.1 Social Relationships were not used when Classifying Bystanders and Subjects in Photos

We developed an automated system to protect bystanders' privacy, who are often strangers to the photographers and or photo-owners. In our attempt to make a general-purpose classifier, we relied only on the visual characteristics of people in the image to classify them as bystanders and subjects. A consequence of this design decision is that our system will consistently flag people as bystanders based on their 'look' even if they are socially related to the photographer/photo-owner. One way to reduce the number of such false positives is to incorporate social relationships into the system; this is possible when the model is incorporated into social networking platforms, where such information is readily available.

8.1.2 Image Obfuscations' Acceptability was Studied from Photo-Viewers' Perspective

We collected images from the internet and applied filters to them. Our study participants viewed these filtered images and rated their utility. Since social media users simultaneously post and view online content, data from our study indicate the likelihood of them adopting the obfuscation when they post photos they own. However, people's preferences for specific obfuscations may differ when they own the images, as they can best judge what information they (do not) want to obscure and how. Future research may investigate what image obfuscations people prefer as *photo-owners* and how that correlates with their preference as viewers.

8.1.3 Unintended Consequences of Image Obfuscations

Obfuscating image regions using filters, such as blurring, pixelating, and masking, may leave visible marks. Such discernible censorships may generate viewers' curiosity and increase their interest

to learn the ‘hidden’ information— a phenomenon known as the ‘Streisand effect’ [215]. Recent advancements in the computer vision algorithms might be a remedy: image regions can be altered to obscure specific information in a non-discriminable way. For example, people’s faces can be replaced with faces created by generative machine learning models. Such fictitious faces look realistic, and thus the alterations are not noticeable anymore, resulting in ‘natural-looking’ obfuscated images. But a concern is that a fictitious face may accidentally match a real person, which might create controversy and have adverse social and professional consequences for that person. Other complications may arise if faces are created with inconsistent or inappropriate properties, e.g., mixing genders or races. One approach to avoid such risks is to just remove information from images without adding any new information. For example, people or other objects in images can be replaced by the image background. But such approaches are not always desirable, since the presence of people and objects partially provide the meanings and contexts of photographs (e.g., a picture of a game without an audience in the gallery would lose part of its meaning). But, it is reasonable to assume that these unintended consequences have a very small chance to occur in practice, and the associated risks are outweighed by the benefits of obfuscations. Further, if using image filters become commonplace in practice, it is unlikely that obfuscated regions of day to day images from an average person would create any undue curiosity. Similarly, the probability of a random face looking similar to a real person is already tiny; on top of that, the contexts in which that person usually appears have to be consistent with the photo context to create any controversy. Nonetheless, future studies may quantify the risks in each of these cases.

8.1.4 Improving the Classification Accuracy of the Bystander Detection Model

Findings from our user study revealed that when humans classify subjects and bystanders in images, they take into account the (dis)similarities in visual appearance and activities among people, as well as the contexts and environments in which those images were taken. But we did not use such

interpersonal relationships and contextual information in the model and classified each person based only on their visual properties. Future work could attempt to infer such information from image data and incorporate them into the decision-making process to improve classification accuracy. One straight forward visual similarity measure between two people could be the correlation between the distributions of RGB values of pixels in the area of the image containing them. Several existing machine learning models (e.g., [124, 177, 227]) that were developed to recognize activities of people in still images can be fine-tuned and used to detect (dis)similarities in activities. Machine learning models for object detection [102] and image captioning [93] can be re-purposed to get a holistic understanding of the image context. All this information can then be plugged in a classification model using, e.g., Conditional Random Fields [116], and used to inferring class assignments of each person.

8.1.5 Designing Better Image Obfuscations

The image obfuscations we designed were combinations of image filters and artistic transforms. But the filters and transforms had distinct visual properties. Their combinations did not result in ‘natural looking’ obfuscations, which was probably the reason for their failure to significantly improve the visual aesthetics of the filtered images. An alternative approach to obscure specific information (e.g., facial expressions) while preserving as much originality of the image as possible might be to utilize generative adversarial deep learning models [73]. Such models can ‘generate’ an image (or portion of an image) and can be constrained to fulfill certain requirements, e.g., change the facial expression but preserve other facial attributes such as the identity or gender of the person [103]. Future work could evaluate the usability of such transforms from image owners, sharers, and viewers’ perspectives.

8.1.6 Visual Interventions to Discourage the Sharing of Privacy-Sensitive Photos

As Amon *et al.* reported, when the participants of their study were asked why they intended to share a meme, some of the most frequently mentioned rationales they provided included *the humorous aspects of the memes* and *how they could relate to the story and/or context of the memes*. The context of a meme is usually set by the added texts and is different from the original context of the photo. This alternative context, which often emphasizes on some characteristics of the photo subject (e.g., an embarrassing activity) or some event depicted in the photo, usually amplifies the photo's humorous aspects. Drawing from the ‘objectification theory’ [64], the person appearing in a meme might often be seen in terms of particular traits, rather than as a human being with complex life and experiences (e.g., emphasizing a person as a “nerd” or a “body-builder”). Thus, the added text shifts the viewers’ attention from people in the photo to their specific peculiarity or some event depicted in the photo. Directing viewers’ attention back to the people in the photo may be an effective way to reverse this process. One way to achieve this might be by manipulating the photo to highlight the subject’s face. Prior research has shown that looking at people’s faces may generate empathy toward them [44], which in turn may inspire more prosocial behaviors and discourage meme-sharing. Research has also found that a familiar face generates more empathy than an unknown face [27]. This inspires another possible intervention: replacing the photo subject’s face with a familiar face (e.g., celebrity) or even the face of the photo-sharer. Both face highlighting and replacement can be done automatically by first detecting the image0region containing the face [96] and then manipulating the pixels in that region either to increase brightness (to highlight) or to replace with other pixel values (e.g., face in-painting [199]). Future research could evaluate the feasibility and efficacy of such interventions.

8.1.7 Evaluating ML Models, Obfuscation Methods, and Behavioral Interventions in the Wild.

As pointed out earlier, our studies did not use photos owned by the participants. As future research, we want to investigate the feasibility of the machine learning models and obfuscation methods that we proposed in a naturalistic environment and using photos taken and/or owned by photo-sharers. This can be done in one of two ways, as described below.

8.1.7.1 A Client-Server Infrastructure to Evaluate Privacy Enhancing Technologies (PETs)

We envision a client-server architecture to study privacy-enhancing technologies (e.g., automated system to detect and obfuscate sensitive image-content, such as bystanders), with ecological validity. The client could be a mobile application that would interact with users and communicate with a server component. When new images are captured, the client would send it to the server for processing, e.g., detecting sensitive information in it and if found, notify the user and offer appropriate obfuscation mechanisms. The server would handle the computation-intensive tasks related to computer vision, image processing, and machine learning. With this infrastructure in place, the proposed models to detect bystanders could be assessed in realistic scenarios. Furthermore, the model can be improved by using additional information collected over time, e.g., the repeated appearance of certain people in photos taken by a particular user would signal that they may not be bystanders. This infrastructure would also allow us to study the acceptance of image obfuscations by the owners of the photos.

8.1.7.2 In-device Processing of Computer Vision Algorithms

One drawback of the client-server architecture is that it would require transferring the visual data from client devices to the server for processing. This may open new attack surfaces for security

and privacy violations. An alternative framework is to perform all the processing on the client side, such as mobile devices. But deep learning models require very high processing power and a large memory space, which are not available in mobile devices. Computer vision and machine learning researchers are inventing ways to reduce the power and memory requirements of such models [51], e.g., by compressing network layers [147]; but such advantages usually come at the expense of reduced accuracy. Future work may assess the feasibility of an in-device framework to automatically detect and obfuscate sensitive image content. As noted earlier, the loss in accuracy due to reduced network size might be compensated by using auxiliary information.

8.1.7.3 Evaluating Behavioral Interventions in the Wild

The test infrastructure described above may also be used to evaluate the proposed and novel interventions to protect bystanders' privacy. As evident from chapter 5, applying obfuscations would inevitably cause a reduction in information, and may also lower the image's visual aesthetics. This may disincentivize photo owners to apply obfuscations to protect bystanders' privacy. The interventions described in chapter 7 may be assessed in this context. Further, this infrastructure would provide a unique opportunity to design and test customized interventions by learning people's personality traits and photo-sharing behaviors using longitudinal data collected by client applications.

CHAPTER 9

Conclusions

Online social networks have become an integral part of people’s daily lives, where they share images depicting daily activities and memorable life events. In addition to the subject matter and primary participants of these events, images capture much incidental information that may be privacy-sensitive, including people who are not relevant for the stories—the ‘bystanders’. Sharing these photos on social media raises numerous privacy concerns for bystanders. Due to large cascaded re-sharing of these images, many of them end up in the public domain, which fuels the building surveillance technologies that have the potential to severely undermine people’s privacy and autonomy. Further, social media users publicly re-share photos as memes, often with demeaning captions that, once they have gone viral, may damage the photo subjects’ social and professional reputation.

This dissertation offers several technical and social solutions to reduce the privacy risks of bystanders and meme-subjects. In Chapter 4, we introduced several machine learning-based models to automatically detect bystanders in images using only image data. Our best model can distinguish bystanders from subjects with high accuracy. It can be easily integrated into social media platforms to protect people’s privacy at scale. In Chapter 5, we presented the assessments of five image filters that have been commonly used to obscure sensitive image content, such as people’s facial expressions. We provide evidence that despite their popularity, the filters failed to protect privacy in most of the cases that we examined. Further, the few effective filters produced filtered images that were not satisfactory to viewers. We addressed the lack of *effective* and *usable* image filters in Chapter 6 by designing new obfuscations through combining privacy filters with artistic transforms. We evaluated these novel obfuscations through a user study that was detailed in Chapter 6. Chapter 7 was focused on understanding how individual differences affect photo-sharing behaviors so that customized interventions can be developed to discourage social media users in creating and

sharing photos (such as memes) that may violate other people's privacy. The contributions of this thesis are essential building blocks towards a comprehensive socio-technical approach to alleviate people's privacy risks in the context of sharing photos on social media.

BIBLIOGRAPHY

- [1] 12.6 - Reducing Structural Multicollinearity.
- [2] 10 Internet memes that ruined lives, 2016.
- [3] Facebook, 2020.
- [4] Flickr, 2020.
- [5] Instagram, 2020.
- [6] Qualtrics. <https://www.qualtrics.com>, 2020.
- [7] SnapChat, 2020.
- [8] WhatsApp, 2020.
- [9] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. *ACM Comput. Surv.*, 50(3), 8 2017.
- [10] Alessandro Acquisti and Ralph Gross. Imagined communities: Awareness, information sharing, and privacy on the {F}acebook. In *Privacy enhancing technologies*, pages 36–58. Springer, 2006.
- [11] Alessandro Acquisti, Ralph Gross, and Frederic D Stutzman. Face recognition and privacy in the age of augmented reality. *Journal of Privacy and Confidentiality*, 6(2):1, 2014.
- [12] Anne Adams, Sally Jo Cunningham, Masood Masoodian, and University of Waikato. Sharing, privacy and trust issues for photo collections. Technical report, 2007.

- [13] Paarijaat Aditya, Rijurekha Sen, Peter Druschel, Seong Joon Oh, Rodrigo Benenson, Mario Fritz, Bernt Schiele, Bobby Bhattacharjee, and Tong Tong Wu. I-Pic: A Platform for Privacy-Compliant Image Capture. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '16, pages 235–248, New York, NY, USA, 2016. ACM.
- [14] Shane Ahern, Dean Eckles, Nathaniel S Good, Simon King, Mor Naaman, and Rahul Nair. Over-exposed?: Privacy Patterns and Considerations in Online and Mobile Photo Sharing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 357–366, New York, NY, USA, 2007. ACM.
- [15] Taslima Akter, Bryan Dosono, Tousif Ahmed, Apu Kapadia, and Bryan Semaan. "i am uncomfortable sharing what i can't see": Privacy concerns of the visually impaired with camera based assistive applications. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1929–1948. USENIX Association, August 2020.
- [16] Mary Jean Amon, Rakibul Hasan, Kurt Hugenberg, Bennett I Bertenthal, and Apu Kapadia. Influencing Photo Sharing Decisions on Social Media: A Case of Paradoxical Findings. In *the Proceedings of the IEEE Symposium on Security & Privacy (SP '20)*. IEEE Computer Society, 5 2020.
- [17] Mark Andrejevic and Neil Selwyn. Facial recognition technology and the end of privacy for good., 2020.
- [18] Denise Anthony, Celeste Campos-Castillo, and Christine Horne. Toward a Sociology of Privacy. *Annual Review of Sociology*, 43(1):249–269, 2017.
- [19] Paritosh Bahirat, Yangyang He, Abhilash Menon, and Bart Knijnenburg. A Data-Driven Approach to Developing IoT Privacy-Setting Interfaces. In *23rd International Conference on*

Intelligent User Interfaces, IUI '18, page 165–176, New York, NY, USA, 2018. Association for Computing Machinery.

- [20] Melissa Bateson, Daniel Nettle, and Gilbert Roberts. Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, 2(3):412–414, 2006.
- [21] BBC. 'Can't hide it forever': The model who became a meme, 2016.
- [22] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and et al. VizWiz: Nearly Real-Time Answers to Visual Questions. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, page 333–342, New York, NY, USA, 2010. Association for Computing Machinery.
- [23] Roberta Biolcati and Stefano Passini. Narcissism and self-esteem: Different motivations for selfie posting behaviors. *Cogent Psychology*, 5(1), 2018.
- [24] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K Nayar. Face Swapping: Automatically Replacing Faces in Photographs. *ACM Trans. Graph.*, 27(3):39:1–39:8, 8 2008.
- [25] YouTube Official Blog. Face blurring: when footage requires anonymity. Blog, 7 2012.
- [26] Cheng Bo, Guobin Shen, Jie Liu, Xiang-Yang Li, YongGuang Zhang, and Feng Zhao. Privacy.Tag: Privacy Concern Expressed and Respected. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*, SenSys '14, pages 163–176, New York, NY, USA, 2014. ACM.
- [27] Stéphane Bouchard, François Bernier, Eric Boivin, Stéphanie Dumoulin, Mylène Laforest, Tanya Guitard, Geneviève Robillard, Johana Monthuy-Blanc, and Patrice Renaud. Empathy

Toward Virtual Humans Depicting a Known or Unknown Person Expressing Pain. *Cyberpsychology, behavior and social networking*, 16:61–71, 2013.

- [28] Adrien Bousseau, Fabrice Neyret, Joëlle Thollot, and David Salesin. Video Watercolorization Using Bidirectional Texture Advection. *ACM Trans. Graph.*, 26(3), 7 2007.
- [29] danah boyd. *Taken out of context: American teen sociality in networked publics*. PhD thesis, University of California, Berkeley, 2008.
- [30] danah boyd. *It's complicated: The social lives of networked teens*. Yale University Press, 2014.
- [31] Michael Boyle, Christopher Edwards, and Saul Greenberg. The Effects of Filtered Video on Awareness and Privacy. In *ACM Conference on Computer Supported Cooperative Work, CSCW '00*, pages 1–10, New York, NY, USA, 2000. ACM.
- [32] Karla Brkic, Ivan Sikiric, Tomislav Hrkac, and Zoran Kalafatic. I Know That Person: Generative Full Body and Face De-Identification of People in Images. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1319–1328. IEEE, 2017.
- [33] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon’s Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.
- [34] By Dawn C. Chmielewski. YouTube, Instagram And Snapchat All More Popular Than Facebook Among Teens, Pew Reports, 2018.
- [35] James P Byrnes, David C Miller, and William D Schafer. Gender differences in risk taking: a meta-analysis. *Psychological bulletin*, 125(3):367, 1999.

- [36] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *arXiv preprint arXiv:1812.08008*, 2018.
- [37] Christopher J Carpenter. Narcissism on Facebook: Self-promotional and anti-social behavior. *Personality and Individual Differences*, 52(4):482–486, 2012.
- [38] Charles S Carver and Michael F Scheier. Self-focusing effects of dispositional self-consciousness, mirror presence, and audience presence. *Journal of Personality and Social Psychology*, 36(3):324, 1978.
- [39] Gary Charness and Uri Gneezy. Strong Evidence for Gender Differences in Risk Taking. *Journal of Economic Behavior & Organization*, 83(1):50–58, 2012.
- [40] Hyerim Cho, Josh Smith, and Jin Ha Lee. Effects of motivation and tool features on online photo-sharing behavior. *Proceedings of the Association for Information Science and Technology*, 56(1):377–380, 2019.
- [41] Eun Kyoung Choe, Sunny Consolvo, Jaeyeon Jung, Beverly Harrison, and Julie A Kientz. Living in a Glass House: A Survey of Private Moments in the Home. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp ’11, pages 41–44, New York, NY, USA, 2011. ACM.
- [42] Jacob Cohen. Statistical power analysis for the social sciences. 1988.
- [43] Jacob Cohen. A power primer. *Psychological bulletin*, 112(1):155, 1992.
- [44] Jonathan Cole. Empathy needs a face. *Journal of Consciousness Studies*, 8(6-7):51–68, 2001.
- [45] Teresa Correa, Amber Willard Hinsley, and Homero Gil [de Zúñiga]. Who interacts on the Web?: The intersection of users’ personality and social media use. *Computers in Human Behavior*, 26(2):247–253, 2010.

- [46] Dianne Cyr, Milena Head, Hector Larios, and Bing Pan. Exploring human images in website design: a multi-method approach. *MIS quarterly*, pages 539–566, 2009.
- [47] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying Aesthetics in Photographic Images Using a Computational Approach. In Aleš Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision – ECCV 2006*, pages 288–301, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.
- [48] Patrick Davison. The language of internet memes. *The social media reader*, pages 120–134, 2012.
- [49] George Days. Seven regular people whose lives got rekt by a meme, 2017.
- [50] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [51] Yunbin Deng. Deep learning on mobile devices: a review. In Sos S Agaian, Vijayan K Asari, and Stephen P DelMarco, editors, *Mobile Multimedia/Image Processing, Security, and Applications 2019*, volume 10993, pages 52–66. International Society for Optics and Photonics, SPIE, 2019.
- [52] Tamara Denning, Zakariya Dehlawi, and Tadayoshi Kohno. In Situ with Bystanders of Augmented Reality Glasses: Perspectives on Recording and Privacy-mediating Technologies. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems, CHI ’14*, pages 2377–2386, New York, NY, USA, 2014. ACM.
- [53] Edward Diener and Mark Wallbom. Effects of self-awareness on antinormative behavior. *Journal of Research in Personality*, 10(1):107–111, 1976.

- [54] Koen A Dijkstra, Joop van der Pligt, Gerben A van Kleef, and José H Kerstholt. Deliberation versus intuition: Global versus local processing in judgment and choice. *Journal of Experimental Social Psychology*, 48(5):1156–1161, 2012.
- [55] Mariella Dimiccoli, Juan Marín, and Edison Thomaz. Mitigating Bystander Privacy Concerns in Egocentric Activity Recognition with Deep Learning and Intentional Image Degradation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–18, 1 2018.
- [56] Amanda du Preez and Elanie Lombard. The role of memes in the construction of Facebook personae. *Communicatio*, 40(3):253–270, 2014.
- [57] Jim Edwards. PLANET SELFIE: We’re Now Posting A Staggering 1.8 Billion Photos Every Day. <http://www.businessinsider.com/were-now-posting-a-staggering-18-billion-photos-to-social-media-every-day-2014-5>, 2014.
- [58] Amir Efrati. Read Congress’s Letter About Google Glass Privacy, 2013.
- [59] Thomas Rhys Evans and Gail Steptoe-Warren. Humor Style Clusters: Exploring Managerial Humor. *International Journal of Business Communication*, 55(4):443–454, 2018.
- [60] Rossella Falanga, Maria Elvira De Caroli, and Elisabetta Sagone. Humor Styles, Self-efficacy and Prosocial Tendencies in Middle Adolescents. *Procedia - Social and Behavioral Sciences*, 127:214–218, 2014.
- [61] Andy Field, Jeremy Miles, and Zoë Field. *Discovering statistics using R*. Sage publications, 2012.
- [62] Graham D Finlayson, Michal Mackiewicz, and Anya Hurlbert. Color correction using root-polynomial regression. *IEEE Transactions on Image Processing*, 24(5):1460–1470, 2015.

- [63] Arturo Flores and Serge Belongie. Removing pedestrians from google street view images. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 53–58. IEEE, 2010.
- [64] Barbara L Fredrickson and Tomi-Ann Roberts. OBJECTIFICATION THEORY. *Psychology of Women Quarterly*, 21(2):173–206, 1997.
- [65] Ronald S Friedman, Ayelet Fishbach, Jens Förster, and Lioba Werth. Attentional Priming Effects on Creativity. *Creativity Research Journal*, 15(2-3):277–286, 2003.
- [66] Graeme Galloway. Individual differences in personal humor styles: Identification of prominent patterns and their associates. *Personality and Individual Differences*, 48(5):563–567, 2010.
- [67] Vaibhav Garg, Sameer Patil, Apu Kapadia, and L Jean Camp. Peer-produced Privacy Protection. In *IEEE International Symposium on Technology and Society ({ISTAS})*, pages 147–154, 6 2013.
- [68] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A Neural Algorithm of Artistic Style. *CoRR*, abs/1508.0, 2015.
- [69] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016.
- [70] Will M Gervais and Ara Norenzayan. Like a camera in the sky? Thinking about God increases public self-awareness and socially desirable responding. *Journal of Experimental Social Psychology*, 48(1):298–302, 2012.
- [71] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual Action Recognition With R*CNN. In *The IEEE International Conference on Computer Vision (ICCV)*, 12 2015.
- [72] Erving Goffman and others. *The presentation of self in everyday life*. Harmondsworth London, 1978.

- [73] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In Z Ghahramani, M Welling, C Cortes, N D Lawrence, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.
- [74] Google Street View. Image Acceptance and Privacy Policies, 2018.
- [75] Robert Gove. Using the elbow method to determine the optimal number of clusters for k-means clustering. *URL: <https://bl.ocks.org/rpgove/0060ff3b656618e9136b>*, 17:19, 2015.
- [76] Ralph Gross, Edoardo Airoldi, Bradley Malin, and Latanya Sweeney. Integrating Utility into Face De-identification. In *International Conference on Privacy Enhancing Technologies, PET'05*, pages 227–242, Berlin, Heidelberg, 2006. Springer-Verlag.
- [77] Ralph Gross, Latanya Sweeney, Jeffrey Cohn, Fernando Torre, and Simon Baker. Protecting Privacy in Video Surveillance. chapter Face De-id, pages 129–146. Springer London, London, 2009.
- [78] L Grundlingh. Memes as speech acts. *Social Semiotics*, 28(2):147–168, 2018.
- [79] Isobel Asher Hamilton. Instagram has avoided Facebook’s trust problem, beating its parent as app of choice for Generation Z, 2019.
- [80] William P Hampes. The Relation Between Humor Styles and Empathy. *Europe’s Journal of Psychology*, 6(3):34–45, 2010.
- [81] Sehee Han, Jinyoung Min, and Heeseok Lee. Antecedents of social presence and gratification of social connection needs in SNS: A study of Twitter users and their mobile and non-mobile usage. *International Journal of Information Management*, 35(4):459–471, 2015.
- [82] Rakibul Hasan, Kurt Bertenthal, Bennett Hugenberg, and Apu Kapadia. Your Photo is so Funny that I don’t Mind Violating Your Privacy by Sharing it: Effects of Individual Humor

- Styles on Online Photo-sharing Behaviors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, 2021. To appear.
- [83] Rakibul Hasan, David Crandall, and Mario Fritz Apu Kapadia. Automatically Detecting Bystanders in Photos to Reduce Privacy Risks. In *2020 IEEE Symposium on Security and Privacy (SP)*, Los Alamitos, CA, USA, 5 2020. IEEE Computer Society.
- [84] Rakibul Hasan, Eman Hassan, Yifang Li, Kelly Caine, David J Crandall, Roberto Hoyle, and Apu Kapadia. Viewer Experience of Obscuring Scene Elements in Photos to Enhance Privacy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 47:1–47:13, New York, NY, USA, 2018. ACM.
- [85] Rakibul Hasan, Yifang Li, Eman Hassan, Kelly Caine, David J. Crandall, Roberto Hoyle, and Apu Kapadia. Can privacy be satisfying? On improving viewer satisfaction for privacy-enhanced photos using aesthetic transforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, volume 14, page 25. ACM, 2019.
- [86] E T Hassan, R Hasan, P Shaffer, D Crandall, and A Kapadia. Cartooning for Enhanced Privacy in Lifelogging and Streaming Videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1333–1342, 7 2017.
- [87] James Hays and Irfan Essa. Image and video based painterly animation. In *International Symposium on Non-photorealistic Animation and Rendering*, pages 113–120, New York, NY, USA, 2004. ACM, ACM.
- [88] Jianping He, Bin Liu, Deguang Kong, Xuan Bao, Na Wang, Hongxia Jin, and George Kesidis. PuPPIeS: Transformation-Supported Personalized Privacy Preserving Partial Image Sharing. In *IEEE International Conference on Dependable Systems and Networks*, Atlanta, Georgia USA, 2014. IEEE Computer Society.

- [89] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [90] Benjamin Henne, Christian Szongott, and Matthew Smith. SnapMe if You Can: Privacy Threats of Other Peoples’ Geo-tagged Media and What We Can Do About It. In *Proceedings of the Sixth ACM Conference on Security and Privacy in Wireless and Mobile Networks*, WiSec ’13, pages 95–106, New York, NY, USA, 2013. ACM.
- [91] Kashmir Hill. The Secretive Company That Might End Privacy as We Know It., 2020.
- [92] Jennifer Hofmann, Tracey Platt, Chloé Lau, and Jorge Torres-Marín. Gender differences in humor-related traits, humor appreciation, production, comprehension,(neural) responses, use, and correlates: A systematic review.
- [93] M D Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Comput. Surv.*, 51(6), 2019.
- [94] Mariea Grubbs Hoy and George Milne. Gender Differences in Privacy-Related Measures for Young Adult Facebook Users. *Journal of Interactive Advertising*, 10(2):28–45, 2010.
- [95] Roberto Hoyle, Robert Templeman, Steven Armes, Denise Anthony, David Crandall, and Apu Kapadia. Privacy Behaviors of Lifeloggers Using Wearable Cameras. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp ’14, pages 571–582, New York, NY, USA, 2014. ACM.
- [96] Peiyun Hu and Deva Ramanan. Finding Tiny Faces. *CoRR*, abs/1612.0, 2016.
- [97] Scott E Hudson and Ian Smith. Techniques for Addressing Fundamental Privacy and Disruption Tradeoffs in Awareness Support Systems. In *Proceedings of the 1996 ACM Conference*

on Computer Supported Cooperative Work, CSCW '96, pages 248–257, New York, NY, USA, 1996. ACM.

- [98] Daniel Hunt and Eric Langstedt. The Influence of Personality Factors and Motives on Photographic Communication. *The Journal of Social Media in Society*, 3(2), 2014.
- [99] Daniel S Hunt, Carolyn A Lin, and David J Atkin. Communicating Social Relationships via the Use of Photo-Messaging. *Journal of Broadcasting & Electronic Media*, 58(2):234–252, 2014.
- [100] Panagiotis Ilia, Iasonas Polakis, Elias Athanasopoulos, Federico Maggi, and Sotiris Ioannidis. Face/off: Preventing privacy leakage from photos in social networks. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 781–792. ACM, 2015.
- [101] Isobel Asher Hamilton. Instagram has avoided Facebook’s trust problem, beating its parent as app of choice for Generation Z, 2019.
- [102] L Jiao, F Zhang, F Liu, S Yang, L Li, Z Feng, and R Qu. A Survey of Deep Learning-Based Object Detection. *IEEE Access*, 7:128837–128868, 2019.
- [103] A Jourabloo, X Yin, and X Liu. Attribute preserved face de-identification. In *2015 International Conference on Biometrics (ICB)*, pages 278–285, 5 2015.
- [104] Sanjay Kairam, Joseph ‘Jofish’ Kaye, John Alexis Guerra-Gomez, and David A Shamma. Snap Decisions?: How Users, Content, and Aesthetics Interact to Shape Photo Sharing Behaviors. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI ’16, pages 113–124, New York, NY, USA, 2016. ACM.

- [105] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara Kiesler. Privacy Attitudes of Mechanical Turk Workers and the U.S. Public. In *10th Symposium On Usable Privacy and Security ({SOUPS} 2014)*, pages 37–49, Menlo Park, CA, 7 2014. {USENIX} Association.
- [106] R M Khan and M A Khan. Academic sojourners, culture shock and intercultural adaptation: A trend analysis. *Studies About Languages*, 10:38–46, 2007.
- [107] Hope King. Wildly popular Prisma app just made a major breakthrough, 2016.
- [108] Kate Knibbs. 1.8 billion images are uploaded every day, 2020.
- [109] Bart P Knijnenburg. Simplifying Privacy Decisions: Towards Interactive and Adaptive Solutions. In *Decisions@ RecSys*, pages 40–41, 2013.
- [110] Mohammed Korayem, Robert Templeman, Dennis Chen, David Crandall, and Apu Kapadia. Enhancing Lifelogging Privacy by Detecting Screens. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI ’16, pages 4309–4314, New York, NY, USA, 2016. ACM.
- [111] P Korshunov and T Ebrahimi. Using face morphing to protect privacy. In *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on*, pages 208–213, Krakow, Poland, 8 2013. IEEE Computer Society.
- [112] Yi-Cheng Ku, Rui Chen, and Han Zhang. Why do users continue using social networking sites? An exploratory study of members in the United States and Taiwan. *Information & Management*, 50(7):571–581, 2013.
- [113] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallochi, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The Open Images Dataset V4. *International Journal of Computer Vision*, 2020.

- [114] Jan Eric Kyprianidis and Jürgen Döllner. Image Abstraction by Structure Adaptive Filtering. In *Proc. EG UK Theory and Practice of Computer Graphics*, pages 51–58, Manchester, UK, 2008. Eurographics Association.
- [115] Ernesto Leon la Rosa-Carrillo. *On the language of Internet Memes*. PhD thesis, 2015.
- [116] John Lafferty, Andrew McCallum, and Fernando C N Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.
- [117] Daniel Lakens. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4:863, 2013.
- [118] H Lee and A Kobsa. Privacy preference modeling and prediction in a simulated campuswide IoT environment. In *2017 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 276–285, 3 2017.
- [119] Anja K. Leist and Daniela Müller. Humor Types Show Different Patterns of Self-Regulation, Self-Esteem, and Well-Being. *Journal of Happiness Studies*, 14(2):551–569, 2013.
- [120] Amanda Lenhart, Kristen Purcell, Aaron Smith, and Kathryn Zickuhr. Social Media & Mobile Internet Use among Teens and Young Adults. *Pew internet & American life project*, 2010.
- [121] A Li, Q Li, and W Gao. PrivacyCamera: Cooperative Privacy-Aware Photographing with Mobile Phones. In *2016 13th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 1–9, 6 2016.
- [122] F Li, Z Sun, A Li, B Niu, H Li, and G Cao. HideMe: Privacy-Preserving Photo Sharing on Social Networks. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, pages 154–162, 4 2019.

- [123] Shan Li, Weihong Deng, and JunPing Du. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593. IEEE, 2017.
- [124] Yang Li, Kan Li, and Xinxin Wang. Recognizing actions in images by fusing multiple body structure cues. *Pattern Recognition*, 104:107341, 2020.
- [125] Yifang Li, Nishant Vishwamitra, Bart P Knijnenburg, Hongxin Hu, and Kelly Caine. Effectiveness and Users’ Experience of Obfuscation as a Privacy-Enhancing Technology for Sharing Photos. *Proceedings of the ACM: Human Computer Interaction (PACM)*, 2018.
- [126] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.
- [127] Bin Liu, Mads Schaaruup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun (Aerin) Zhang, Norman Sadeh, Yuvraj Agarwal, and Alessandro Acquisti. Follow My Recommendations: A Personalized Privacy Assistant for Mobile App Permissions. In *Twelfth Symposium on Usable Privacy and Security ({SOUPS} 2016)*, pages 27–41, Denver, CO, 6 2016. {USENIX} Association.
- [128] Di Liu, Randolph G Bias, Matthew Lease, and Rebecca Kuipers. Crowdsourcing for usability testing. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–10, 2012.
- [129] Natasha Lomas. Teens favoring Snapchat and Instagram over Facebook -says e-marketer, 2017.

- [130] Aqdas Malik, Amandeep Dhir, and Marko Nieminen. Uses and Gratifications of digital photo sharing on Facebook. *Telematics and Informatics*, 33(1):129–138, 2016.
- [131] Farhad Manjoo. Why Instagram Is Becoming Facebook’s Next Facebook), 2017.
- [132] Rod A Martin, Jessica M Lastuk, Jennifer Jeffery, Philip A Vernon, and Livia Veselka. Relationships between the Dark Triad and humor styles: A replication and extension. *Personality and Individual Differences*, 52(2):178–182, 2012.
- [133] Rod A Martin, Patricia Puhlik-Doris, Gwen Larsen, Jeanette Gray, and Kelly Weir. Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. *Journal of Research in Personality*, 37(1):48–75, 2003.
- [134] By Craig McCarthy and Aaron Feis. Rogue NYPD cops are using facial recognition app Clearview., 2020.
- [135] Michael P McCreery and S [Kathleen Krach]. How the human is the catalyst: Personality, aggressive fantasy, and proactive-reactive aggression among users of social media. *Personality and Individual Differences*, 133:91–95, 2018.
- [136] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. Defeating Image Obfuscation with Deep Learning. *CoRR*, abs/1609.0, 2016.
- [137] Adam W Meade and S Bartholomew Craig. Identifying careless responses in survey data. *Psychological methods*, 17(3):437–455, 9 2012.
- [138] Thelwall Mike and Vis Farida. Gender and image sharing on Facebook, Twitter, Instagram, Snapchat and WhatsApp in the UK: Hobbying alone or filtering for friends? *Aslib Journal of Information Management*, 69(6):702–720, 1 2017.
- [139] Ian Miller and Gerald Cupchik. Meme creation and sharing processes: individuals shaping the masses. *arXiv preprint arXiv:1406.7579*, 2014.

- [140] Ryan M Milner. *The world made meme: Discourse and identity in participatory media*. PhD thesis, University of Kansas, 2012.
- [141] V Mirjalili and A Ross. Soft biometric privacy: Retaining biometric utility of face images while perturbing gender. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 564–573, 2017.
- [142] G Misra and J M Such. PACMAN: Personal Agent for Access Control in Social Media. *IEEE Internet Computing*, 21(6):18–26, 2017.
- [143] Kelly Moore and James C McElroy. The influence of personality on Facebook usage, wall postings, and regret. *Computers in Human Behavior*, 28(1):267–274, 2012.
- [144] Carol Moser. *Impulse Buying: Designing for Self-Control with E-commerce*. PhD thesis, 2020.
- [145] Vivian Genaro Motti and Kelly Caine. Users' Privacy Concerns About Wearables. In Michael Brenner, Nicolas Christin, Benjamin Johnson, and Kurt Rohloff, editors, *Financial Cryptography and Data Security*, pages 231–244, Berlin, Heidelberg, 2015. Springer Berlin Heidelberg.
- [146] Y Nakashima, T Koyama, N Yokoya, and N Babaguchi. Facial expression preserving privacy protection using image melding. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, pages 1–6, Torino, Italy, 6 2015. IEEE Computer Society.
- [147] K Nan, S Liu, J Du, and H Liu. Deep model compression for mobile platforms: A survey. *Tsinghua Science and Technology*, 24(6):677–693, 2019.
- [148] Carman Neustaedter, Saul Greenberg, and Michael Boyle. Blur Filtration Fails to Preserve Privacy for Home-based Video Conferencing. *ACM Trans. Comput.-Hum. Interact.*, 13(1):1–36, 3 2006.

- [149] M Nishiyama, T Okabe, I Sato, and Y Sato. Aesthetic quality classification of photographs based on color harmony. In *CVPR 2011*, pages 33–40, 6 2011.
- [150] Helen Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009.
- [151] Anne Oeldorf-Hirsch and S Shyam Sundar. Social and Technological Motivations for Online Photo Sharing. *Journal of Broadcasting and Electronic Media*, 60(4):624–642, 2016.
- [152] Office of the Privacy Commissioner of Canada. Data protection authorities urge Google to address Google Glass concerns, 2013.
- [153] Eileen Y L Ong, Rebecca P Ang, Jim C M Ho, Joylynn C Y Lim, Dion H Goh, Chei Sian Lee, and Alton Y K Chua. Narcissism, extraversion and adolescents’ self-presentation on Facebook. *Personality and Individual Differences*, 50(2):180–185, 2011.
- [154] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2014.
- [155] Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2018.
- [156] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-Octob, pages 3706–3715, 10 2017.
- [157] Jason W Osborne, Anna B Costello, and J Thomas Kellow. Best practices in exploratory factor analysis. *Best practices in quantitative methods*, pages 86–99, 2008.

- [158] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [159] A J Perez, S Zeadally, and S Griffith. Bystanders’ Privacy. *IT Professional*, 19(3):61–65, 2017.
- [160] Pew Research Center. Photo and Video Sharing Grow Online. *Pew Research Center*, 2013.
- [161] Mark R Phillips, Bradley D McAuliff, Margaret Bull Kovera, and Brian L Cutler. Double-blind photoarray administration as a safeguard against investigator bias. *Journal of Applied Psychology*, 84(6):940, 1999.
- [162] Abdul Qodir, Ahmad Muhammad Diponegoro, and Triantoro Safaria. Cyberbullying, happiness, and style of humor among perpetrators: is there a relationship? *Humanities & Social Sciences Reviews*, 7(3):200–206, 2019.
- [163] Lizhen Qu, Gabriela Ferraro, Liyuan Zhou, Weiwei Hou, and Timothy Baldwin. Named entity recognition for novel types by transfer learning. *arXiv preprint arXiv:1610.09914*, 2016.
- [164] M Ra, S Lee, E Miluzzo, and E Zavesky. Do Not Capture: Automated Obscurity for Pervasive Imaging. *IEEE Internet Computing*, 21(3):82–87, 5 2017.
- [165] Moo-Ryong Ra, Ramesh Govindan, and Antonio Ortega. P3: Toward Privacy-preserving Photo Sharing. In *USENIX Conference on Networked Systems Design and Implementation, nsdi’13*, pages 515–528, Berkeley, CA, USA, 2013. USENIX Association.
- [166] Beatrice Rammstedt and Oliver P John. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41(1):203–212, 2007.

- [167] Christina L. Rash and Sally M. Gainsbury. Disconnect between intentions and outcomes: A comparison of regretted text and photo social networking site posts. *Human Behavior and Emerging Technologies*, 1(3):229–239, 2019.
- [168] Yasmeen Rashidi, Tousif Ahmed, Felicia Patel, Emily Fath, Apu Kapadia, Christena Nippert-Eng, and Norman Makoto Su. "You don't want to be the next meme": College Students' Workarounds to Manage Privacy in the Era of Pervasive Photography. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 143–157, Baltimore, MD, 2018. USENIX Association.
- [169] Yasmeen Rashidi, Apu Kapadia, Christena Nippert-Eng, and Norman Makoto Su. "it's easier than causing confrontation": Sanctioning strategies to maintain social norms of content sharing and privacy on social media. *Proceedings of the ACM Journal: Human-Computer Interaction: Computer Supported Cooperative Work and Social Computing (CSCW '20)*, 4(CSCW1):23:1–23:25, May 2020.
- [170] Yasmeen Rashidi, Apu Kapadia, Christena Nippert-Eng, and Norman Makoto Su. "It's easier than causing confrontation": Sanctioning Strategies to Maintain Social Norms of Content Sharing and Privacy on Social Media. *To appear in the Proceedings of the ACM Journal: Human-Computer Interaction: Computer Supported Cooperative Work and Social Computing (CSCW '20)*, 2020.
- [171] E M Redmiles, S Kross, and M L Mazurek. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *2019 IEEE Symposium on Security and Privacy (SP)*, volume 00, pages 227–244.
- [172] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In C Cortes, N D Lawrence, D D Lee,

- M Sugiyama, and R Garnett, editors, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 39, pages 1137–1149. Curran Associates, Inc., 2017.
- [173] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to Anonymize Faces for Privacy Preserving Action Detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 9 2018.
- [174] Joel Ross, Lilly Irani, M Six Silberman, Andrew Zaldivar, and Bill Tomlinson. Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, pages 2863–2872, New York, NY, USA, 2010. ACM.
- [175] Mike Rugnetta. Are Memes & Internet Culture Creating a Singularity?, 2012.
- [176] Tracii Ryan and Sophia Xenos. Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage. *Computers in Human Behavior*, 27(5):1658–1664, 2011.
- [177] M Safaei and H Foroosh. Still Image Action Recognition by Predicting Spatial-Temporal Pixel Evolution. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 111–120, 2019.
- [178] Carlos Salavera, Pablo Usán, and Laurane Jarie. Styles of humor and social skills in students. Gender differences. *Current Psychology*, 39(2):571–580, 2020.
- [179] Peter Seddon and Min-Yen Kiew. A Partial Test and Development of Delone and Mclean's Model of IS Success. *Australasian Journal of Information Systems*, 4(1), 1996.
- [180] SHANE LARKIN. MEMES THAT DESTROYED LIVES, 2017.
- [181] Ryan Shaw. Recognition markets and visual privacy. *UnBlinking: New Perspectives on Visual Privacy in the 21st Century*, 2006.

- [182] Kim [Bartel Sheehan]. An investigation of gender differences in on-line privacy concerns and resultant behaviors. *Journal of Interactive Marketing*, 13(4):24–38, 1999.
- [183] Kim Bartel Sheehan. Toward a Typology of Internet Users and Online Privacy Concerns. *The Information Society*, 18(1):21–32, 2002.
- [184] Jiayu Shu, Rui Zheng, and Pan Hui. Cardea: Context-Aware Visual Privacy Protection from Pervasive Cameras. *arXiv preprint arXiv:1610.00889*, 2016.
- [185] Jiayu Shu, Rui Zheng, and Pan Hui. Your Privacy Is in Your Hand: Interactive Visual Privacy Control with Tags and Gestures. In Nishanth Sastry and Sandip Chakraborty, editors, *Communication Systems and Networks*, pages 24–43. Springer International Publishing, Cham, 2017.
- [186] Andra Siibak. Constructing the self through the photo selection-visual impression management on social networking websites. *Cyberpsychology: Journal of psychosocial research on cyberspace*, 3(1), 2009.
- [187] Paul J Silvia and T Shelley Duval. Objective Self-Awareness Theory: Recent Progress and Enduring Problems. *Personality and Social Psychology Review*, 5(3):230–241, 2001.
- [188] T Sim and L Zhang. Controllable Face Privacy. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 04, pages 1–8, 5 2015.
- [189] Samarth Singhal, Carman Neustaedter, Thecla Schiphorst, Anthony Tang, Abhisekh Patra, and Rui Pan. You Are Being Watched: Bystanders’ Perspective on the Use of Camera Devices in Public Spaces. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA ’16, pages 3197–3203, New York, NY, USA, 2016. ACM.

- [190] Manya Sleeper, Rebecca Balebako, Sauvik Das, Amber Lynn McConahy, Jason Wiese, and Lorrie Faith Cranor. The Post That Wasn'T: Exploring Self-censorship on Facebook. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 793–802, New York, NY, USA, 2013. ACM.
- [191] Manya Sleeper, Justin Cranshaw, Patrick Gage Kelley, Blase Ur, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. "I Read My Twitter the Next Morning and Was Astonished": A Conversational Perspective on Twitter Regrets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 3277–3286, New York, NY, USA, 2013. ACM.
- [192] Kit Smith. 53 Incredible Facebook Statistics and Facts, 2019.
- [193] A C Squicciarini, A Novelli, D Lin, C Caragea, and H Zhong. From Tag to Protect: A Tag-Driven Policy Recommender System for Image Sharing. In *2017 15th Annual Conference on Privacy, Security and Trust (PST)*, pages 337–33709, 2017.
- [194] Michelle Starr. Facial recognition app matches strangers to online profiles, 2014.
- [195] Julian Steil, Marion Koelle, Wilko Heuten, Susanne Boll, and Andreas Bulling. PrivacEye: Privacy-Preserving First-Person Vision Using Image Features and Eye Movement Analysis. *arXiv preprint arXiv:1801.04457*, 2018.
- [196] Charles Steinfield, Nicole B Ellison, and Cliff Lampe. Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of Applied Developmental Psychology*, 29(6):434–445, 2008.
- [197] J M Such and N Criado. Resolving Multi-Party Privacy Conflicts in Social Media. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1851–1863, 2016.

- [198] Jose M Such, Joel Porter, Sören Preibusch, and Adam Joinson. Photo Privacy Conflicts in Social Media: A Large-scale Empirical Study. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 3821–3832, New York, NY, USA, 2017. ACM.
- [199] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and Effective Obfuscation by Head Inpainting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2018.
- [200] Yongjun Sung, Jung-Ah Lee, Eunice Kim, and Sejung Marina Choi. Why we post selfies: Understanding motivations for posting pictures of oneself. *Personality and Individual Differences*, 97:260–265, 2016.
- [201] Robert Templeman, Mohammed Korayem, David Cr, and Apu Kapadia. PlaceAvoider: Steering first-person cameras away from sensitive spaces. In *In NDSS*, 2014.
- [202] Sigal Tifferet. Gender differences in privacy tendencies on social network sites: A meta-analysis. *Computers in Human Behavior*, 93:1–12, 2019.
- [203] Victor Tiscareno, Kevin Johnson, and Cindy Lawrence. Systems and Methods for Receiving Infrared Data with a Camera Designed to Detect Images based on Visible Light, 2011.
- [204] Andrew R Todd, Galen V Bodenhausen, Jennifer A Richeson, and Adam D Galinsky. Perspective taking combats automatic expressions of racial bias. *Journal of Personality and Social Psychology*, 100(6):1027, 2011.
- [205] C Tomasi and R Manduchi. Bilateral filtering for gray and color images. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 839–846, 1 1998.
- [206] Leman Pinar Tosun. Motives for Facebook use and expressing “true self” on the Internet. *Computers in Human Behavior*, 28(4):1510–1517, 2012.

- [207] Lam Tran, Deguang Kong, Hongxia Jin, and Ji Liu. Privacy-CNH: A Framework to Detect Photo Privacy with Convolutional Neural Network Using Hierarchical Features. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 1317–1323. AAAI Press, 2016.
- [208] Khai N Truong, Shwetak N Patel, Jay W Summet, and Gregory D Abowd. Preventing Camera Recording by Designing a Capture-Resistant Environment. In Michael Beigl, Stephen Intille, Jun Rekimoto, and Hideyuki Tokuda, editors, *UbiComp 2005: Ubiquitous Computing*, pages 73–86, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.
- [209] Amrisha Vaish, Malinda Carpenter, and Michael Tomasello. Sympathy through affective perspective taking and its relation to prosocial behavior in toddlers. *Developmental psychology*, 45(2):534, 2009.
- [210] Livia Veselka, Julie Aitken Schermer, Rod A Martin, and Philip A Vernon. Relations between humor styles and the Dark Triad traits of personality. *Personality and Individual Differences*, 48(6):772–774, 2010.
- [211] Nishant Vishwamitra, Yifang Li, Kevin Wang, Hongxin Hu, Kelly Caine, and Gail-Joon Ahn. Towards PII-based Multiparty Access Control for Photo Sharing in Online Social Networks. In *Proceedings of the 22nd ACM on Symposium on Access Control Models and Technologies*, pages 155–166. ACM, 2017.
- [212] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, and Alexandra Kunz. User experiences of TORPEDO: TOoltip-poweRed Phishing Email DetectiOn. *Computers & Security*, 71:100–113, 2017.
- [213] Eric B. Weiser. #Me: Narcissism and its facets as predictors of selfie-posting frequency. *Personality and Individual Differences*, 86:477–481, 2015.

- [214] Wikipedia. List of memory biases.
- [215] Wikipedia contributors. Streisand effect.
- [216] Pamela Wisniewski, Heather Lipford, and David Wilson. Fighting for My Space: Coping Mechanisms for sns Boundary Regulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 609–618, New York, NY, USA, 2012. ACM.
- [217] Pamela J Wisniewski, Bart P Knijnenburg, and Heather Richter Lipford. Making privacy personal: Profiling social network users to inform privacy education and nudging. *International Journal of Human-Computer Studies*, 98:95–108, 2017.
- [218] H Xu, M Reddy, and X L Zhang. Collaborative privacy practices in social media. In *Proc. CSCW*, 2011.
- [219] Jeremy A Yip and Rod A Martin. Sense of humor, emotional intelligence, and social competence. *Journal of Research in Personality*, 40(6):1202–1208, 2006.
- [220] An Gie Yong and Sean Pearce. A beginner’s guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology*, 9(2):79–94, 2013.
- [221] Virgil Zeigler-Hill and Avi Besser. Humor style mediates the association between pathological narcissism and self-esteem. *Personality and Individual Differences*, 50(8):1196–1201, 2011.
- [222] Eva-Maria Zeissig, Chantal Lidynia, Luisa Vervier, Andera Gadeib, and Martina Ziefle. Online Privacy Perceptions of Older Adults. In Jia Zhou and Gavriel Salvendy, editors, *Human Aspects of IT for the Aged Population. Applications, Services and Contexts*, pages 181–200, Cham, 2017. Springer International Publishing.
- [223] Hang Zhang and Kristin J Dana. Multi-style Generative Network for Real-time Transfer. *CoRR*, abs/1703.0, 2017.

- [224] L Zhang, T Jung, C Liu, X Ding, X Y Li, and Y Liu. POP: Privacy-Preserving Outsourced Photo Sharing and Searching for Mobile Devices. In *Distributed Computing Systems (ICDCS), 2015 IEEE 35th International Conference on*, pages 308–317, Columbus, Ohio, USA, 6 2015. IEEE Computer Society.
- [225] Lan Zhang, Kebin Liu, Xiang-Yang Li, Cihang Liu, Xuan Ding, and Yunhao Liu. Privacy-friendly Photo Capturing and Sharing System. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp ’16*, pages 524–534, New York, NY, USA, 2016. ACM.
- [226] Yin Zhang, Leo Shing-Tung Tang, and Louis Leung. Gratifications, Collective Self-Esteem, Online Emotional Openness, and Traitlike Communication Apprehension as Predictors of Facebook Uses. *Cyberpsychology, Behavior, and Social Networking*, 14(12):733–739, 2011.
- [227] Zhichen Zhao, Huimin Ma, and Shaodi You. Single Image Action Recognition Using Semantic Body Part Actions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 10 2017.
- [228] Tomasz Zukowski and Irwin Brown. Examining the Influence of Demographic Factors on Internet Users’ Information Privacy Concerns. In *Proceedings of the 2007 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries, SAICSIT ’07*, page 197–204, New York, NY, USA, 2007. Association for Computing Machinery.

APPENDIX A

Additional analyses for chapter 4

A.1 Predictive power of each feature

In section 4.4.2, we saw that the features are associated with the classification rationales (Table 4.3 and Table 4.4). Next, we want to investigate how effectively the features can distinguish between *subject* and *bystander*. Results of logistic regression analyses using each of the features individually as predictors are reported in Table A.1. The χ^2 statistic indicates how well the data fit the model, where higher values indicate better fit. The value of the R^2 statistic refers to the amount of variance of the outcome variable that was explained by the predictor variable. Note that *Replaceable* has the largest values for both of the statistics, which is intuitive since it is almost a synonym for *being a bystander*. For each predictor, the *Odds ratio* with 95% confidence interval is also presented in Table A.1. *Odds ratio* refers to the effect of increasing a predictor's variable by one unit to the outcome variable in a multiplicative scale. For example, increasing the value for *Pose* by one unit will *increase* the odds of a person of being classified as a *subject* by 4.48 times than before. On the other hand, increasing the value for *Replaceable* by one unit will *decrease* the odds of a person of being classified as a *subject* by 11.11 times than before. When used as individual predictors, the features *Replaceable*, *Awareness*, *Willingness*, *Pose*, and *Comfort* all have reasonably high effects on the outcome variable and the data fit the model well enough. But *Photo place* is not a very effective predictor ($OR=1.41$, $\chi^2=101.6$). The *Size* feature has large effect on the outcome, but using this as an individual predictor it may be noisy as suggested by the lower χ^2 value.

A.2 Correlation among pairs of features

Table A.2 shows Pearson's product moment correlation coefficients (r) between pairs of features. Almost all pairs of features have medium to high correlations between them [42]. In particular,

Table A.1: Effectiveness of visual features used individually as predictors to classify *subject* and *bystander*. All χ^2 statistics are significant at $p < 0.0001$ level.

Predictor	Odds ratio	[2.5%]	97.5%]	χ^2	R^2
Replaceable	0.09	0.07	0.10	2254.41	0.44
Awareness	5.19	4.66	5.78	1476.37	0.29
Willingness	4.38	3.96	4.86	1247.30	0.24
Pose	4.48	4.01	5.00	1146.42	0.22
Comfort	4.05	3.66	4.48	1121.78	0.22
Size	5.23	4.52	6.05	960.15	0.19
Distance	0.31	0.29	0.34	930.95	0.18
Number of people	0.50	0.46	0.54	410.43	0.08
Photographer intention	0.53	0.49	0.57	330.39	0.06
Photo place	1.41	1.32	1.51	101.60	0.02

Table A.2: Correlation coefficients (r) between pairs of visual features. Each coefficient is significant at $p < .001$ level.

Feature1	Feature2	r	Feature1	Feature2	r
Awareness	Pose	0.88	Pose	Comfort	0.73
	Comfort	0.75		Willingness	0.76
	Willingness	0.79		Replaceable	-0.48
	Replaceable	-0.57		Size	0.42
	Size	0.45		Distance	-0.34
	Distance	-0.37			
Comfort	Willingness	0.86	Willingness	Replaceable	-0.52
	Replaceable	-0.49		Size	0.39
	Size	0.37		Distance	-0.33
	Distance	-0.32			
Replaceable	Size	-0.44	Size	Distance	-0.48
	Distance	0.42		Number of people	-0.43
	Number of people	0.31			

Awareness is highly correlated with most of the other features, suggesting that they collectively contain the same information as the ‘*Awareness*’ feature.

Table A.3 shows the VIF for each feature before and after removing the highly correlated ‘*Awareness*’ feature.

Table A.3: Variance inflation factor (VIF) of predictor variables when all predictors were used (Initial VIF) and after *Awareness* was removed (Updated VIF).

Variable	Initial VIF	Updated VIF
Awareness	5.80	-
Pose	4.67	2.62
Comfort	4.24	4.23
Willingness	5.01	4.72
Photographer intention	1.11	1.1
Replaceable	1.77	1.73
Photo place	1.14	1.13
Size	1.71	1.7
Distance	1.42	1.42
Number of people	1.27	1.27

A.3 Predicting *high-level concepts* from the *proxy* features

As detailed in the Section 4.3.3.3, we infer the *high-level concepts* using the proxy features – human related features, body-pose features, and emotion – using linear regression models. For each of the *high-level concepts*, the mean and standard deviations for training loss, *mean squared error (MSE)*, and *mean absolute error (MAE)* across a 10-fold cross-validation of the regression models are shown in Table A.4. The error values are interpreted in relation to the range of scores of the outcome variable, since the same error score would indicate a good or bad model depending on whether the range is large or small, respectively. In our case, all the concepts except *Willingness* have the same range of possible values (-3 to 3), and so the prediction errors for them can be compared. *Photographer’s intention* has the highest loss and prediction errors. This was expected given that it is more nuanced than the other concepts, and highly depends on the overall context of the image and interactions among people in it. Since we only used features from the cropped portion of the image containing a single person for prediction, the loss and errors go higher. On average *Comfort* could be predicted with the highest accuracy. All the other concepts have about the same losses and prediction errors. Finally, *Willingness* has a smaller range of possible values (-2 to 2), and accordingly, smaller loss and error values.

Table A.4: Results of predicting *high-level concepts* using image data. Columns show means and standard deviations of *loss*, *mean absolute error (MAE)*, and *mean squared error (MSE)* of a 10-fold cross-validation.

Outcome	Loss		MAE		MSE	
	Mean	SD	Mean	SD	Mean	SD
Awareness	1.79	0.07	1.04	0.02	1.65	0.06
Photographer’s intention	2.65	0.15	1.30	0.04	2.47	0.15
Replaceable	1.60	0.08	0.98	0.03	1.46	0.07
Pose	1.99	0.14	1.08	0.05	1.81	0.14
Comfort	0.81	0.05	0.67	0.03	0.72	0.05
Willingness	0.45	0.02	0.50	0.02	0.40	0.02

Table A.5: Percentage of participants agreed with the final classification label and number of photos with that agreement values.

Agreement	Number of photos
33%	256
50%	208
67%	1308
75%	300
100%	1309

A.4 Agreement among the annotators

Table A.5 presents the percentages of agreement among the study participants and the number of images for each percentage. We included percentages for which the number of photos are greater than 100.

A.5 Comparing with human annotators

Figure A.1 shows Receiver Operating Characteristic (ROC) plots for classifiers trained and tested on images with 67% and 100% agreements among the survey participants.

A.6 Attention check questions

The two images shown in Fig. A.2 were used for attention check questions. We asked **Which of the following statements is true for the person inside the green rectangle in the photo?**

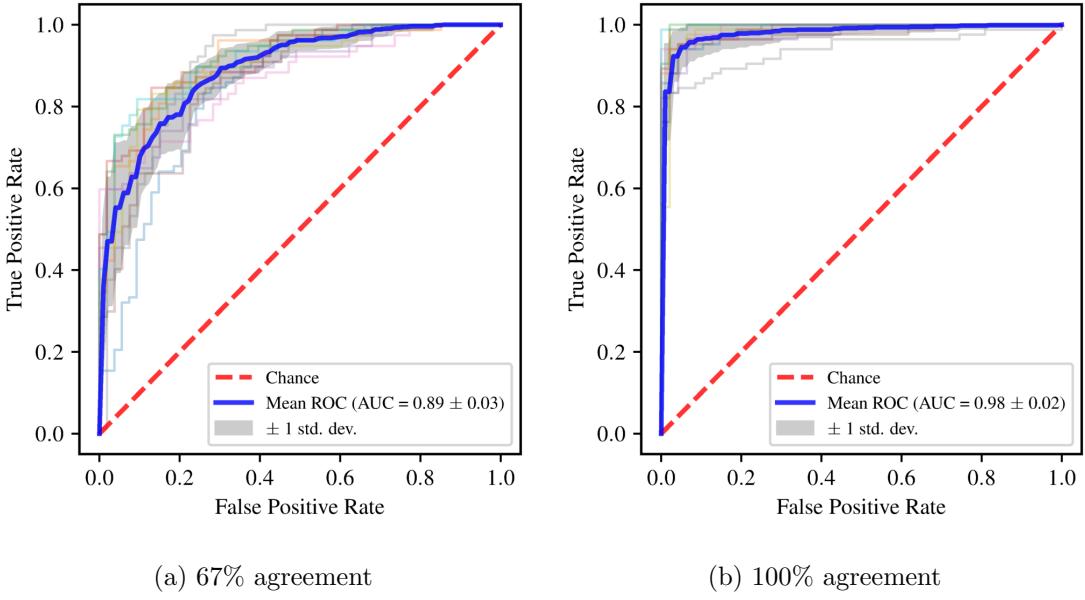


Figure A.1: Receiver operating characteristic (ROC) plots for classifiers trained and tested on images with (a) 67% agreement and (b) 100% agreement among the survey participants.

with answer options i) There is a person with some of the major body parts visible (such as face, head, torso); ii) There is a person but with no major body part visible (e.g., only hands or feet are visible); iii) There is just a depiction/representation of a person but not a real person (e.g., a poster/photo/sculpture of a person); iv) There is something else inside the box; and v) I don't see any box. Since the persons in the bounding boxes are clearly visible, if any survey participant responded with any option other than the first one, we marked it as wrong.

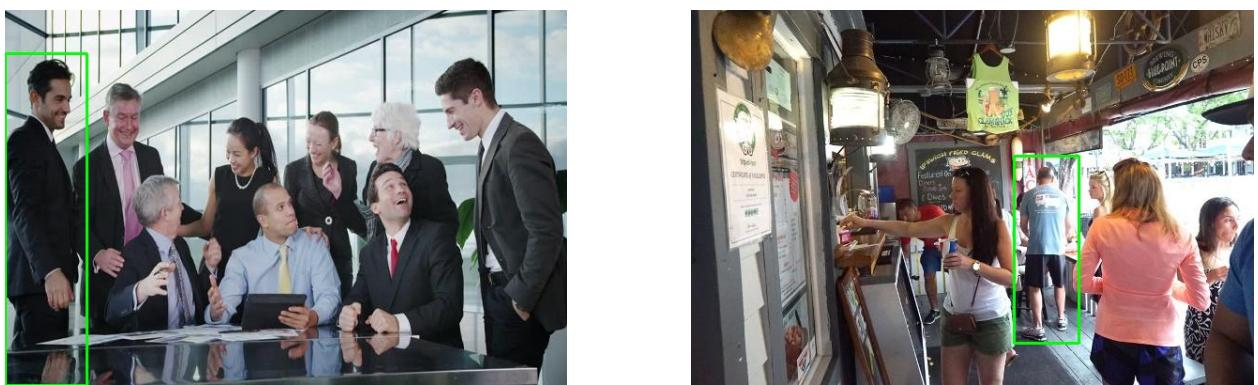


Figure A.2: Images used for attention check questions.

APPENDIX B

Additional methodological details for Chapter 5

B.1 Selecting Filter Levels

We conducted a separate user study to select the three levels for *blur*, *pixel*, and *edge* filters.

We built an application that would load a set of images in random order and apply one of the three filters in predefined regions (see Figure B.1). These filtered images are then displayed in the graphical user interface (GUI) of this application. A slider allowed one to change the filter value.

At first the filters were applied with their highest values possible (50 for *blur* and *pixel*, 1 for *edge*).

The filter value was then reduced in proportion to the displacement of the slider, and the filter was reapplied with its new value.

We collected data from ten volunteers, who were shown filtered images in random order, and were asked two questions pertaining to the filtered regions:

1. The first question sought to determine at what filter level high-level details of the image became apparent to the participant. For example, for *computer monitor* we asked *What is the object inside the filtered region?*; for *indoor environment* we asked *Was this photo taken indoors or outside?*.
2. The second question sought to determine the filter level at which lower level details became apparent. For example, for *computer monitor* we asked *What is the application shown in the monitor inside the filtered region?*; for *indoor environment* we asked *What type of indoor place is shown in the image?*.

These two questions were asked always in the order presented here (because otherwise revealing low level details first would also reveal higher level details). At the beginning of the study the first filtered image with the maximum value was displayed and the first question was asked. If the

participant could not answer it, he/she was instructed to move the slider towards right to reduce the filter value and make the filtered region clearer until he/she could answer the question with confidence. Once the participant could answer the first question correctly, we recorded the current filter value as the ‘high’ value. Likewise we used the second question to determine the ‘low’ value. This process was repeated for all other images for each participant.

Once we gathered data from all ten participants, we calculated the high, mid, and low level filter values for our study as follows:

- **High level:** Average of the high values + 1 standard deviation of high values.
- **Medium level:** Average of the high values.
- **Low level:** Average of the low values + 1 standard deviation of low values.

Except silhouette, for all the other filters we used Matlab programming language to apply them over images in predefined rectangular regions. The *blur* filter is applied using circular averaging filter, with the radius as an input parameter that can be tuned by the user using the slider (see Figure B.1). For *pixel* filter, the image was divided into grids with size $S \times S$, which is also a parameter tuned by the user, where each grid’s RGB values are replaced by the average value in the transformed image. In both cases each slider tick represents changes in the tuned parameter value by one, and the change is used to produce the newly transformed image. The maximum value of the transformation is 50 which is believed to provide complete obfuscation with respect to the image size. Finally the *edge* filter is applied using the ‘canny’ edge detector. With a tuned parameter that represents a canny hysteresis high threshold value t , the low threshold value is $0.4 \times t$ which is the default setting in Matlab. Each slider tick value represents a change in resolution of 0.01; the allowed parameter values are $\in (0, 1)$

Since *silhouette* preserves the object shape, we manually drew *silhouettes* on objects. We applied the filters in such a way that they cover only the object/attribute of interest in each image. For

example, to cover facial expressions, we apply the filter only on the face whereas to cover dress, we apply the filter on the entire body. Please note that we did not use *silhouette* for the following scenarios: indoor and outdoor environment (both general and specific) and text (both on document and on screen).

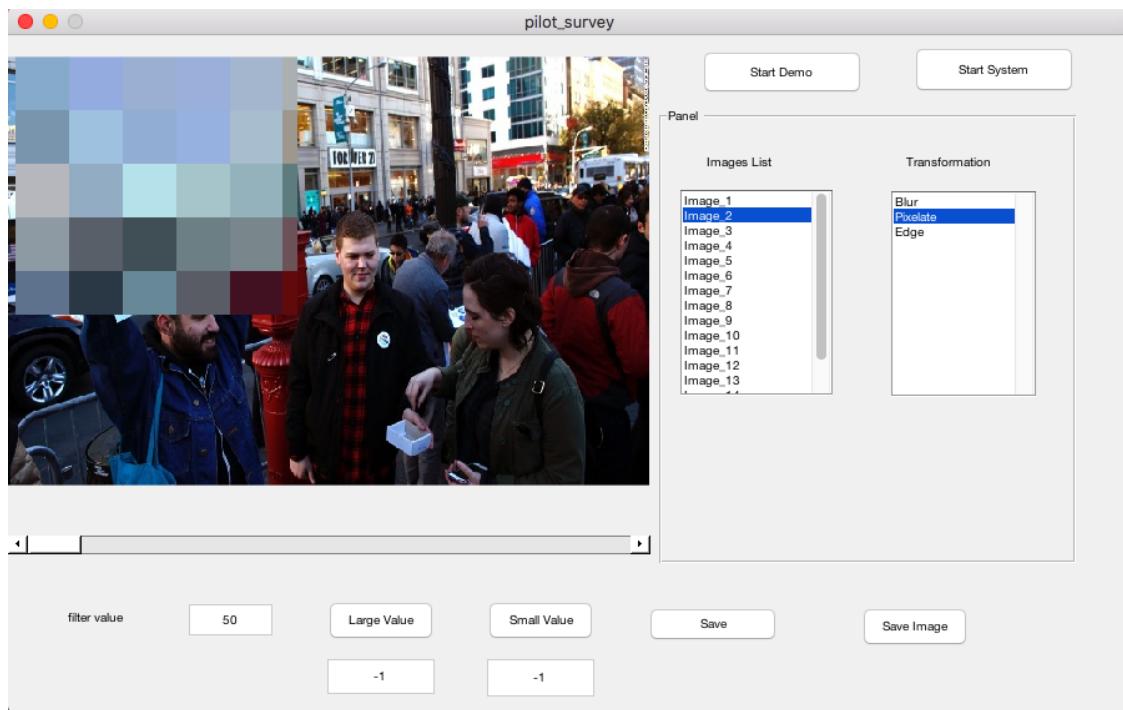


Figure B.1: Screenshot of the application we developed to select filter levels.

APPENDIX C

Survey for Chapter 7

C.1 Questionnaire

Social Media Usage Questionnaire

- Which social media platforms do you have an account for? (Select all that apply.)
 1. Facebook, 2. Instagram 3. Pinterest 4. Snapchat 5. Twitter 6. Myspace 7. Flicker 8. Other
(Please describe)
- How often you visit social media?
 1. Never, 2. Less than once in a month, 3. Once in a month, 4. Multiple times in a month, 5. Once in a week, 6. Multiple times in a week, 7. Once in a day, 8. Multiple times in a day
- What social media platform do you use to share photos online the most? (Select all that apply.)
 1. Facebook, 2. Instagram 3. Pinterest 4. Snapchat 5. Twitter 6. Myspace 7. Flicker 8. Other
(Please describe)
- When you share photos online, who do you typically share them with?
 1. Friends/connections, 2. General viewers/public, 3. Both
- How often do you share photos on social media?
 1. Never, 2. Less than once in a month, 3. Once in a month, 4. Multiple times in a month, 5. Once in a week, 6. Multiple times in a week, 7. Once in a day, 8. Multiple times in a day
- How often do you share pictures taken by you, your friends, or your family on social media?
 1. Never, 2. Less than once in a month, 3. Once in a month, 4. Multiple times in a month, 5. Once in a week, 6. Multiple times in a week, 7. Once in a day, 8. Multiple times in a day
- How often do you share pictures on social media that you found on the internet or that other people took (not including your friends, family or other people you personally know.)? 1. Never, 2. Less than once in a month, 3. Once in a month, 4. Multiple times in a month, 5. Once in a week, 6. Multiple times in a week, 7. Once in a day, 8. Multiple times in a day

Experimental Manipulation

- (Baseline condition) How likely are you to share this photo on social media?
1.Extremely unlikely, 2. Moderately unlikely, 3. Slightly unlikely, 4. Neither unlikely nor likely, 5. Slightly likely, 6. Moderately likely, 7. Extremely likely
- (Perspective taking condition) If this was a photo of you, how likely are you to share this photo on social media?
1.Extremely unlikely, 2. Moderately unlikely, 3. Slightly unlikely, 4. Neither unlikely nor likely, 5. Slightly likely, 6. Moderately likely, 7. Extremely likely
- (Privacy perspective condition) Taking into account the privacy of the person in the photo, how likely are you to share this photo on social media?
1.Extremely unlikely, 2. Moderately unlikely, 3. Slightly unlikely, 4. Neither unlikely nor likely, 5. Slightly likely, 6. Moderately likely, 7. Extremely likely

Social Media Privacy Questionnaire: Answer each of the questions below with options: i) Yes
ii) Maybe iii) No

1. Has anyone ever shared a picture of you online that you did not want them to share?
2. Has anyone ever shared a picture of you online that you felt violated your privacy?
3. Have you ever been embarrassed by a picture of yourself that has been posted online?
4. Have you ever regretted posting a picture of yourself online?
5. Have you ever accidentally posted a picture of yourself online that you did not want to share?
6. Have you ever shared an embarrassing picture online of someone else you know?
7. Have you ever regretted posting a picture online of someone else you know?

8. Have you ever posted a picture online of someone else you know, which may have violated his or her privacy?
9. Have you ever shared an embarrassing picture online of a stranger (someone that you do not personally know)?
10. Have you ever regretted posting a picture online of a stranger (i.e., someone you do not personally know)?
11. Have you ever posted a picture of a stranger (i.e., someone you do not personally know), which may have violated his or her privacy?
12. Do people you know post pictures that might be embarrassing to other people?
13. Has anyone you know regretted posting a picture of another person?
14. Has anyone you know regretted posting a picture of themselves?
15. Has anyone you know posted a picture that may have violated someone's privacy?

Privacy Preference Question

Are you a private person who keeps to yourself or an open person who enjoys sharing with others?

- 1) Very private ... 7) Very open

Rakibul Hasan: Curriculum Vitae

✉ rakhsan@iu.edu | 🏠 rakib062.github.io | 🎓 Google Scholar

Education

- December 2020 Ph.D. in *Computer Science* (Minor: Cognitive Science)
Indiana University Bloomington, USA.
- April 2012 Bachelor of Science in *Computer Science and Engineering.*
Bangladesh University of Engineering and Technology, Dhaka, Bangladesh.

Awards and Honors

APRIL, 2020	Student grant to attend and present my research at the <i>Doctoral Consortium, CHI 2020.</i>
AUGUST, 2019	USENIX Association grant to attend the 15th Symposium, on Usable Privacy and Security, SOUPS 2019
JULY, 2019	NSF travel grant for attending <i>Privacy Enhancing Technologies Symposium (PETS 2019)</i>
APRIL, 2019	Best poster award in <i>Midwest Security Workshop (MSW 2019)</i>
JULY, 2018	NSF travel grant for attending <i>Privacy Enhancing Technologies Symposium (PETS 2018)</i>
JULY, 2017	NSF travel grant for attending <i>The Bright and Dark Sides of Computer Vision: Challenges and Opportunities for Privacy and Security (CV COPS 2017)</i>

Publications

Refereed conference and journal

1. **Rakibul Hasan**, Bennett I. Bertenthal, Kurt Hugenberg, Apu Kapadia, “Your Photo is so Funny that I don’t Mind Violating Your Privacy by Sharing it: Effects of Individual Humor Styles on Online Photo-sharing Behaviors.” To appear at 2021 ACM CHI Conference on Human Factors in Computing Systems (**CHI’21**).

2.  **Rakibul Hasan**, David Crandall, Mario Fritz, Apu Kapadia, “Automatically Detecting Bystanders in Photos to Reduce Privacy Risks”. *IEEE Symposium on Security and Privacy, 2020 (Oakland'20, 12.3% acceptance rate)* (Runner up at the 2020 CNIL-Inria Privacy Protection Award).
3. Mary Jean Amon, **Rakibul Hasan**, Kurt Hugenberg, Bennett Bertenthal, and Apu Kapadia “Influencing Photo Sharing Decisions on Social Media: A Case of Paradoxical Findings.” *IEEE Symposium on Security and Privacy, 2020 (Oakland'20, 12.3% acceptance rate)*.
4. **Rakibul Hasan**, Yifang Li, Eman Hassan, Kelly Caine, David Crandal, Roberto Hoyle, and Apu Kapadia, “Can Privacy Be Satisfying? On Improving Viewer Satisfaction for Privacy-Enhanced Photos Using Aesthetic Transforms”. ACM CHI Conference on Human Factors in Computing Systems (**CHI'19**, 23.8% acceptance rate).
5. **Rakibul Hasan**, Eman Hassan, Yifang Li, Kelly Caine, David Crandal, Roberto Hoyle, and Apu Kapadia, “Viewer Experience of Obscuring Scene Elements in Photos to Enhance Privacy”. ACM CHI Conference on Human Factors in Computing Systems (**CHI'18**, 25.7% acceptance rate).
6. M. R. Islam, S. Rahaman, **Rakibul Hasan**, R. R. Noel, A. Salekin and H. S. Ferdous, “A Novel Approach for Constructing Emulator for Microsoft Kinect XBOX 360 Sensor in the .NET Platform,” 4th International Conference on Intelligent Systems, Modelling and Simulation, 2012.
7. R. R. Noel, A. Salekin, R. Islam, S. Rahaman, **Rakibul Hasan**, and H. S. Ferdous. “A natural user interface classroom based on Kinect”. In IEEE Learning Technology Newsletter, volume 13, October 2011.

Workshops

1. Tousif Ahmed, **Rakibul Hasan**, Kay Connelly, David Crandall, and Apu Kapadia “Conveying Situational Information to People with Visual Impairments”. Workshop at ACM CHI Conference on Human Factors in Computing Systems. (**CHI'19**)
2. Eman Hassan, **Rakibul Hasan**, Patrick Shaffer, David Crandall, and Apu Kapadia, “Cartooning for Enhanced Privacy in Lifelogging and Streaming Video,”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop on Computer Vision Challenges and Opportunities for Privacy and Security (**CVPRW 2017**).

Extended abstracts | Posters | Short talks

1. “Your Photo is so Funny that I don’t Mind Violating Your Privacy by Sharing it: Individual Humor Styles and Photo-sharing Behaviors”. Extended abstract at Symposium on Usable Privacy and Security (**SOUPS 2020**).
2. “Influencing Photo-sharing Behaviors on Social Media to Reduce Privacy Risks”. Short talk at IEEE Symposium on Security and Privacy, 2020 (**S&P 2020**).
3. “Reducing Privacy Risks in the Context of Sharing Photos Online”. Extended abstract at the Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020 (**CHI’20**)
4. “Automatically Detecting Bystanders in Images to Reduce Privacy Risks”. Open Day for Privacy, Usability, and Transparency (**PUT@PETs’19**)
5.  “Learning to Detect Bystanders in Images”. Midwest Security Workshop (Best poster award at **MSW’19**)

Professional Experience

CURRENT Postdoctoral Researcher at *CISPA Helmholtz Center for Information Security, Germany.*

SUMMER 2019 Research Intern at *International Computer Science Institute, Berkeley, CA.*

SUMMER 2018 Research Intern at *Max Planck Institute for Informatics, Germany.*

FALL 2015 - FALL 2020 Research assistant at *Indiana University.*

SEPT 2013 - JULY 2015 Software Engineer in *R&D Sports Team, VizRT.*

APR 2012 - AUG 2013 Software Engineer at *Reve Systems.*

Volunteer Services

- PC member of CVPR workshp: The Bright and Dark Sides of Computer Vision (CV-COPS’2019)

- Served as a student volunteer for CHI 2020 PC meetings.
- Served as an external reviewer for ICWSM'18, ICWSM'19, CHI'20, CHI'21, CSCW'20, CSCW'21, CVCOPS'18, CVCOPS'19.
- Reviewed articles (on request) submitted to *USENIX Security Symposium 2020* and *ACM Transactions on the Web Journal*