IS470 MID TERM REPORT
IDENTITY OBFUSCATION IN PERSON IMAGES

Student: Muhammad Naufal (ID: 01319703)
Supervisor: Prof. Sun Qianru

ABSTRACT

The increasing usage of social media means that everyday, around 1.8 billion photos are uploaded onto the internet [6]. With varying degrees of protection from each social media platform, this means that the user is faced with varying degrees of threat to their privacy. The user would also have no means of knowing if their privacy has been compromised, as there is no efficient method to trace every photo on the internet [5]. As such, there represents a need for AI systems to obfuscate images where private information is shown. Private information may range from face images, location images, as well as artifact images. However as majority of private information lies in our face images, that would be the main information that we would not want to compromise [1]. Once our privacy has been compromised, we are vulnerable to attacks by bad actors, such as scammers [19]. In order to prevent such an outcome, many methods have been introduced to obfuscate face images [29; 14]. They range from blurring faces, swirling, black-out, in-painting as well as masking [24]. This is still a much worked on topic in computer vision as many novel techniques have been introduced recently [27]. They range from using face in-painting as well as face de-identification [25; 14]. We propose a 3-stage framework for face image obfuscation, that leverages on using attribute manipulation, face landmarks and image context to generate new face images for purposes of obfuscation. We also introduce a novel method of word embedding on face attributes to extract information in higher dimensions.

1 INTRODUCTION

1.1 PROBLEM STATEMENT

The rise of social media and information sharing has lead to increasing concerns and questions on privacy. Combined with advancements in the field of machine learning, this has lead to many privacy questions being unanswered. Armed with web-scapers at scale, and state-of-the-art person recognizers, companies such as 'Clearview AI' are able to identify a person with just a single face image [9]. Although such companies have claimed that the usage is mainly used to track down criminals, there is no guarantee that the technology would fall into the hands of bad actors. In order to combat this, social media platforms allows users some degree of privacy protection by allowing users to set preferences for visibility of their photos. This can be done by setting filters to their pictures to blur faces or by purposely hiding the images from certain audiences [15]. However this approach has become increasingly ineffective due to the recent advances in machine learning to enable person recognition [20].

1.2 OBJECTIVES AND GOALS OF RESEARCH

We aim to explore new methods of obfuscation in person images. Particularly those related to conditional image generation. In our case, we want to explore the impact of face image generation conditional on face attributes, face landmarks and image context. Disentangling the image into these three factors will be done in stage-1.

We also aim to improve on the quality of face image generation, as well as to explore novel techniques on the conditional image generation process. Such as the impact of word embedding on image reconstruction. Image reconstruction from the three factors will be done in stage-2.

Finally we aim to answer the question, would face attribute manipulation and word embedding help in generating better quality face images for the purposes of identity obfuscation. The image generation task will be done in stage-3.

1.3 RELATED WORKS

1.3.1 SEMI SUPERVISED IMAGE RECONSTRUCTION

Our work in image reconstruction is closely related to that of [17] where the authors proposed a novel technique of person image generation via sampling disentangled factors of a person image. While the focus of the research was on the task of person re-identification via an entirely self-supervised pipeline, certain elements of their pipeline have been adapted to achieve our objectives. They disentangled the person images into pose, background and foreground, whereas we disentangle the face images into face landmarks, image context and face attributes. Hence the difference is that due to the self-supervise nature of their method, they were not able to extract image attributes. We do this by training a multi-label face attribute classifier (FAC), that can extract face attributes at test time. Further, as they intended to manipulate all three factors, they trained separate mapping functions for each factor, in comparison, our approach does not require a mapping function for each factor as the aim is to only manipulate face attributes, keeping the other two factors fixed.

1.3.2 CONDITIONAL IMAGE GENERATION ON FACE ATTRIBUTES

In [14] the authors aimed to achieve a higher level understanding of face image obfuscation, such that they were able to ask specific questions surrounding the privacy of an image. They introduced the Privacy-Preserving Attribute Selection (PPAS) algorithm to select and update face attributes, followed by image generation via style transfer. While they propose using attribute manipulation as an initial method before image generation, they do not take into account the relation between attributes, such as how close one attribute is to another in a vector space. They do however take into account the distribution of attributes which is more of a frequency based approach rather than a relationship based approach. We hope that in doing so, we will be able to reconstruct better quality images as we are able to provide face attribute information in higher dimensions, rather than just doing one-hot encoding of attributes, which do not give us much information. Further for the FAC task, the authors proposed using 40 different random forest classifiers whereas we propose using just a single multi-label classifier to extract face attributes. As mentioned in [18], such methods have higher cost in terms of run-time.

1.3.3 MULTI-LABEL FACE ATTRIBUTE CLASSIFICATION

The authors in [18] introduce a novel method for FAC, where they incorporate multi-task learning and multi-label classification for the FAC task. In order to classify images they split the learning task into two stages, one with shared features and one with task-specific features. In comparison to our approach we do not adopt multi-task as the objective of our research is to disentangle the face attributes from the face landmarks, hence we opt for two distinct tasks instead, landmark detection and FAC. Further the authors proposed grouping attributes into 2 classes, subjective and objective. Our approach makes no distinction between attributes as we adopt a multi-label approach that treats each label as distinct, hence we do not take into account correlations between attributes. Relations between attributes will only be accounted for at stage-2, the image reconstruction task.

1.4 RESEARCH MOTIVATION

Many methods have been introduced for effective image obfuscation as many novel techniques have been introduced recently [27]. The problem is also general in the sense that there does not appear to be a one-size fits all solution. There have been successful attempts at face swapping to obfuscate identities [2], however at the same time there have been successful attempts at detecting swapped faces [4]. Hence, there is a need for a variety of high-quality methods for obfuscation.

2 METHODOLOGY

2.1 OVERVIEW OF ALL 3-STAGES

Our goal is to generate realistic face images by manipulating face attributes. In order to manipulate the face attributes, we need to disentangle the face attributes from the face image first, followed by image reconstruction and then face attribute manipulation. We do this via a 3 stage framework, where training and testing is done in stage-1 and stage-2, and only testing is done in stage-3. In stage-1, we disentangle the face image into three parts, the face outline via landmark detection [11], the image context via face-masking and the face attributes via a face attribute classifier (FAC) with a ResNet-50 architecture [8]. In stage-2, we will focus on the task of image reconstruction from the three parts mentioned in stage-1. We do this in a self-supervised manner, via a decoder with a U-Net architecture [22]. Before passing the three parts into the decoder, we pass the 40 face attributes through a word2vec encoder [21], to generate word embeddings. This word embeddings allow us to increase the dimensionality of our earlier attributes via training a word2vec neural network and then extracting the weights of the network. This will in turn allow us to extract more information from the face attributes. In the final stage, using the trained decoder in stage 2, we will generate new faces by manipulating the face attribute combination of the input image. Stage-3 is our testing stage as want to see if the images generated would perform well against the state-of-the-art machine recognizers as well as human recognizers.

2.2 STAGE 1

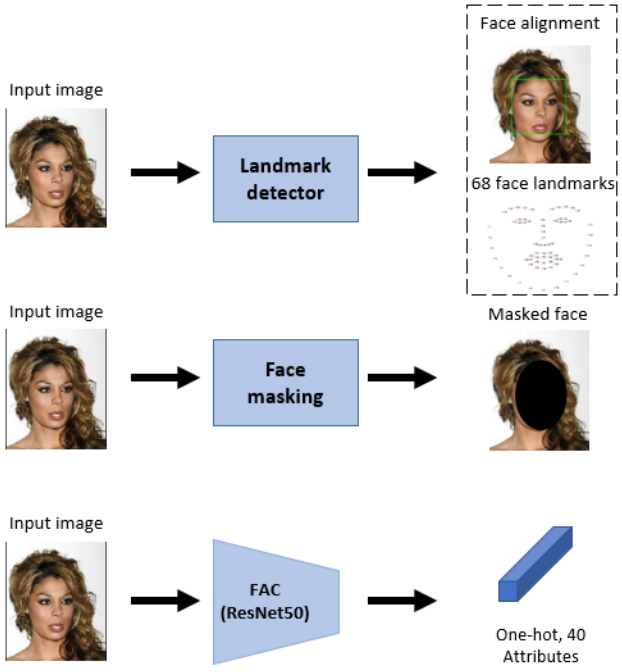


Figure 1: Overview of stage 1, landmark detection, face masking, FAC

2.2.1 LANDMARK DETECTION

For the landmark detection task, we follow the methods introduced in [23] using a pre-trained model available via dlib [12]. The model is able to track the 68 face landmarks in the form of *(x,y)* coordinates. These 68 points will be used in the next step which is face masking.

2.2.2 FACE MASKING

For face masking we use the ellipse method provided by cv2 [3]. Using the face landmarks detected earlier, we will draw a black ellipse over the face region, such that only the image context is left.

2.2.3 FACE ATTRIBUTE ESTIMATION

For face attribute classification. We adopt the ResNet50 architecture as mentioned in [18]. We remove the final output layer and add a flatten layer before adding a final dense layer as our output layer. The final dense layer will have 40 units, corresponding to the 40 attributes we want to classify. For each unit in the output layer, we use sigmoid as the activation function. This will allow us to have 40 separate probabilities, each corresponding to one attribute. The solver we use is adam with default settings [13]. For the loss function, we adopt the binary cross entropy loss as in [7]. The loss function for each sample and each class is formulated as such:

$$L_{FAC}^{j} = -\sum_{i=1}^{2} t_i log(s_i) = -t_1 log(s_1) - (1 - t_1) log(1 - s_1) \tag{1}$$

Where *j* corresponds to the attribute or class and *i* corresponds to the label of ground truth. As our output is binary, t_1 is either 1 or 0, and we have substituted $t_2 = 1 - t_1$ in (1). s_1 is the output of our sigmoid activation, where $0 < s_1 < 1$, and we have substituted $s_2 = 1 - s_1$ in (1). We then extend this loss function to formulate the total loss function at each pass through the network as such:

$$L_{FAC} = \sum_{j=1}^{40} L_{FAC}^{j} \tag{2}$$

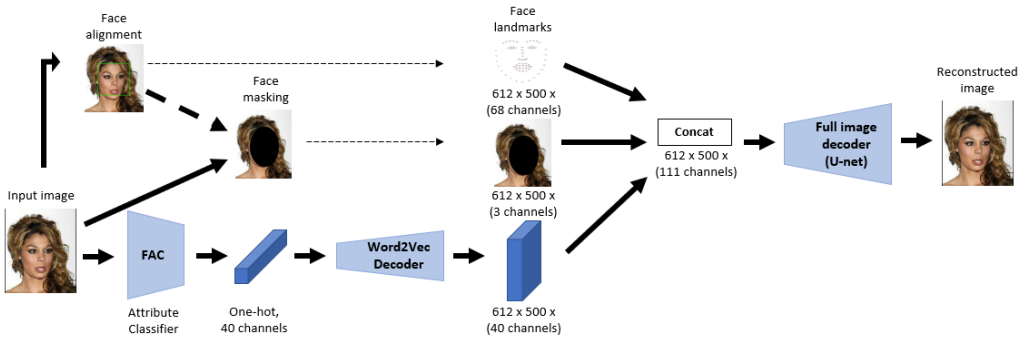


Figure 2: Overview of stage 2, image reconstruction

2.3 STAGE 2

2.3.1 IMAGE RECONSTRUCTION

For image reconstruction, we will adopt the method found in [17] where a self-supervised method was used to reconstruct person images for the task of re-id, that is to augment images to create more images for model training purposes. Figure 2 gives the overview of the intended pipeline. Using the output from stage 1 and with further processing on the output of our FAC, we will have three different blocks to concat before passing it through a decoder. The three blocks are the face landmarks, the face mask, as well as 40 face attributes. For the 40 face attributes, instead of using a one-hot encoding for each image, in relation to what labels are present and which are not, we pass the 40 face attributes through a word2vec encoder [21], to generate word embeddings. This word embeddings allow us to increase the dimensionality of our earlier attributes via training a word2vec neural network and then extracting the weights of the network. This will in turn allow us to extract more information from the face attributes. We use a continuous bag-of-word architecture(CBOW), where the model predicts the current word from surrounding context words. We chose CBOW as the order of context words does not influence prediction in the bag-of-words assumption [28]. After generating the word embeddings, we will concat the 40 channels. For channels where attributes are not present, we will use a null matrix to represent them. This is then followed by concatenating the three different blocks before passing the tensor into the decoder. We will then begin the image reconstruction process. As in [17], we will use both L1 and adversarial loss to optimize the face image. As mentioned by the authors, this combination would result in sharper and more realistic images.

2.4 STAGE 3

2.4.1 FACE GENERATION

For the task of face generation we will use the decoder trained in stage-2. At test time, we will pass the image through the same pipeline as in figure 2, however before the final concat operation, we will manipulate the face attributes. At this stage, the techniques to be used to manipulate the face attributes can either be random or it can follow an algorithm such as PPAS.

3 EXPERIMENTS

3.1 EXPERIMENT SETUP

For the setup of the experiment[1], we used the CelebFaces Attributes Dataset (CelebA) [16]. The data consists of 202,599 face image, with 10,177 identities and 40 binary attributes annotations per image. For training purposes, only a subset of the dataset for use. We use $subset = 0.3$ for our experiments. So for each train,val,test split from the original dataset, we take 30% of each of these splits for our experiments due to memory constraints.

3.2 OUTCOME OF EXPERIMENTS

3.2.1 FACE ATTRIBUTE CLASSIFICATION

For the hyperparameters of our model we set it as such, batch size = 64, number of epochs = 15. For our baseline model, the optimizer that we chose is the Adam with the default learning rates. We set our activation functions in our output layer as sigmoid activations such that we can get the probabilities of each class. The corresponding loss function that we will use for each node is the binary cross entropy. For evaluation criterion, the authors in [18] used the approach of average accuracy over the labels, we will adopt this approach. The experimental results are shown in table 1. We managed to get an average test accuracy of 89.07% of our baseline classifier which is a little off from the top accuracy that the authors managed to achieve. However owing to the fact that we did not manage to train on the entire dataset, it would not be wise to compare with the performance of their experiment. We also trained another classifier however this time, we set the learning rate to a constant rate of 0.0001. The average accuracy for this setting however dropped to 87.9%. The hardware we used is 1x Nvidia RTX 2080ti and the total training time is 53 mins for our baseline and 50 mins for our model with adjusted learning rate. Software used is keras with tensorflow backend.

Table 1: Face attribute classification results

Setting	**Average Accuracy**	**Runtime**
adam	89.07%	53 mins
adam with lr=0.0001	87.90%	50 mins

3.2.2 FACE RECONSTRUCTION AND FACE GENERATION

We are currently in the training stage for the face reconstruction task as well as the face generation task, and do not have experimental results at the moment.

3.2.3 EVALUATION

We will evaluate the generated images using both human and machine assessors.

Human Assessor

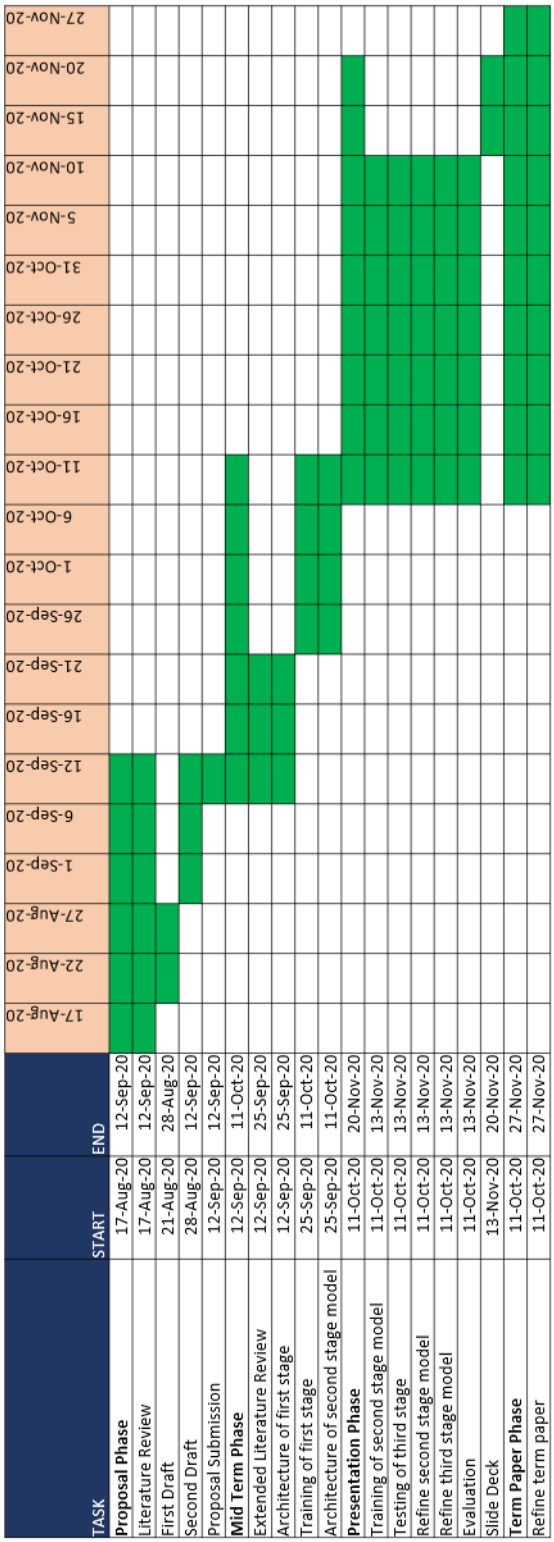
For human assessment, we will pair the original identity (non obfuscated-image) with an image of the same identity with obfuscation. We will then include another 8 images with different identities (non obfuscated-image). We will then instruct the human assessor to pick a pair of images with similar identity. If the human assessor is able to pick up the pair, then it would indicate that the obfuscation is not effective.

Machine Assessor

For machine assessor, as in [26], we will use the state-of-the-art for person recognizer to evaluate our obfuscation method [10].

[1] https://github.com/Juznauf/IS470-public

4 TIMELINE OF WORK



TASK	START	END
Proposal Phase	17-Aug-20	12-Sep-20
Literature Review	17-Aug-20	12-Sep-20
First Draft	21-Aug-20	28-Aug-20
Second Draft	28-Aug-20	12-Sep-20
Proposal Submission	12-Sep-20	12-Sep-20
Mid Term Phase	12-Sep-20	11-Oct-20
Extended Literature Review	12-Sep-20	25-Sep-20
Architecture of first stage	12-Sep-20	25-Sep-20
Training of first stage	25-Sep-20	11-Oct-20
Architecture of second stage model	25-Sep-20	11-Oct-20
Presentation Phase	11-Oct-20	20-Nov-20
Training of second stage model	11-Oct-20	13-Nov-20
Testing of third stage	11-Oct-20	13-Nov-20
Refine second stage model	11-Oct-20	13-Nov-20
Refine third stage model	11-Oct-20	13-Nov-20
Evaluation	11-Oct-20	13-Nov-20
Slide Deck	13-Nov-20	20-Nov-20
Term Paper Phase	11-Oct-20	27-Nov-20
Refine term paper	11-Oct-20	27-Nov-20

REFERENCES

[1] Abhijit Ahaskar. How posting photos online can compromise privacy, Mar 2019. URL https://www.livemint.com/technology/tech-news/how-posting-photos-online-can-compromise-privacy-1553787545932.html.

[2] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K Nayar. Face swapping: automatically replacing faces in photographs. In *ACM SIGGRAPH 2008 papers*, pp. 1–8. 2008.

[3] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[4] Xinyi Ding, Zohreh Raziei, Eric C Larson, Eli V Olinick, Paul Krueger, and Michael Hahsler. Swapped face detection using deep learning and subjective assessment. *EURASIP Journal on Information Security*, 2020:1–12, 2020.

[5] Mary Eising. How to find stolen photos on the internet 2019 guideline. URL https://www.copytrack.com/how-to-find-stolen-images/.

[6] Rose Eveleth. How many photographs of you are out there in the world?, Nov 2015. URL https://www.theatlantic.com/technology/archive/2015/11/how-many-photographs-of-you-are-out-there-in-the-world/413389/#:~:text=In2014,accordingtoMary,intotal150yearsago.

[7] Raul Gomez. Understanding categorical cross-entropy loss, binary cross-entropy loss, softmax loss, logistic loss, focal loss and all those confusing names, 2018. URL https://gombru.github.io/2018/05/23/cross_entropy_loss/.

[8] 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*, pp. 770–778, 2016.

[9] Kashmir Hill. The secretive company that might end privacy as we know it, Jan 2020. URL https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

[10] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Person recognition in personal photo collections. In *Proceedings of the IEEE international conference on computer vision*, pp. 3862–3870, 2015.

[11] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1867–1874, 2014.

[12] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[14] Tao Li and Lei Lin. Anonymousnet: Natural face de-identification with measurable privacy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[15] Yifang Li, Nishant Vishwamitra, Hongxin Hu, Bart P Knijnenburg, and Kelly Caine. Effectiveness and users' experience of face blurring as a privacy protection for sharing photos via online social networks. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 61, pp. 803–807. SAGE Publications Sage CA: Los Angeles, CA, 2017.

[16] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[17] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *The IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[18] Longbiao Mao, Yan Yan, Jing-Hao Xue, and Hanzi Wang. Deep multi-task multi-label cnn for effective facial attribute classification, 2020.

[19] USC Suzanne Dworak-Peck School of Social Work staff. Stalking in the age of social media, Feb 2018. URL https://news.usc.edu/135757/stalking-in-the-age-of-social-media/.

[20] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2016.

[21] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[23] Adrian Rosebrock. Facial landmarks with dlib, opencv, and python, Apr 2020. URL https://www.pyimagesearch.com/2017/04/03/facial-landmarks-dlib-opencv-python/.

[24] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pp. 1528–1540, 2016.

[25] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5050–5059, 2018.

[26] Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. A hybrid model for identity obfuscation by face replacement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 553–569, 2018.

[27] J. Tekli, B. al Bouna, R. Couturier, G. Tekli, Z. al Zein, and M. Kamradt. A framework for evaluating image obfuscation under deep learning-assisted privacy attacks. In *2019 17th International Conference on Privacy, Security and Trust (PST)*, pp. 1–10, Los Alamitos, CA, USA, aug 2019. IEEE Computer Society. doi: 10.1109/PST47121.2019.8949040. URL https://doi.ieeecomputersociety.org/10.1109/PST47121.2019.8949040.

[28] Wikipedia contributors. Word2vec — Wikipedia, the free encyclopedia, 2020. URL https://en.wikipedia.org/w/index.php?title=Word2vec&oldid=982637235. [Online; accessed 11-October-2020].

[29] Xiao Yang, Yinpeng Dong, Tianyu Pang, Jun Zhu, and Hang Su. Towards privacy protection by generating adversarial identity masks. *arXiv preprint arXiv:2003.06814*, 2020.