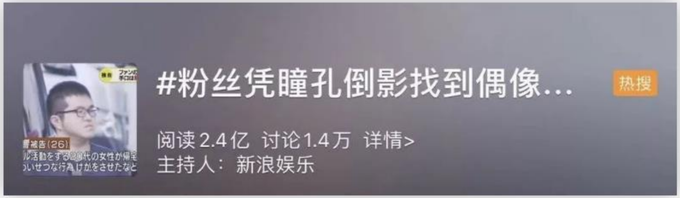
数据隐私方法伦理和实践
Methodology, Ethics and Practice of Data Privacy

6. 隐私检测
Privacy Detection

张兰
中国科学技术大学 计算机学院
2020春季

Privacy risks is horrible but hard to detect.

Ignored details



Additional metadata

Content correlation

Aim to answer:

Is the data private?

&

Why?

Consensus of Privacy Detection

Privacy Related Factors $\xrightarrow{\text{Lead to}}$ Private Data

Could be :

- Metadata
- Specific parts of data
- Specific connections between parts of data
- …

Methods should be effective and interpretable

Overview of Privacy Detection Techniques

» User-defined methods

- Define private related factors by user

Privacy Related Factors → (Lead to) → Private Data

Define factors by user → Detect private data

» Machine Learning Based Methods

- Discover private related factors in a data driven way

Privacy Related Factors → (Lead to) → Private Data

Discover Potential factors by data analysis → Detect private data

Privacy Detection Techniques

Take image, a representative modal of unstructured data, as an example:

- User-defined Methods
  - Image-level user survey
  - Object-level user survey

- Machine Learning Based Method
  - Content sensitiveness
  - Multimodal fusion

Privacy Detection Techniques

Take image, a representative modal of unstructured data, as an example:

- User-defined Methods
  - Image-level user survey
  - Object-level user survey
- Machine Learning Based Method
  - Content sensitiveness
  - Multimodal fusion

User-defined Methods – Image Level

- PicAlert!: A System for Privacy-Aware Image Classification and Retrieval [1]
    - Aim to search related images with key word/Predict privacy label of a new image
    - Main work
        - Label data manually (image-level)
        - Extract features and train a classifier

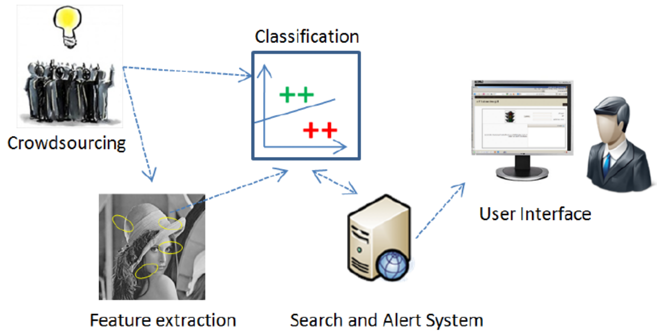


Figure 1: System architecture overview.

User-defined Methods – Image Level

- An image-level privacy dataset
  - Obtain image privacy dataset with label by crowdsourcing
    - 37535 images from Flickr
    - 81 judgers between 10 and 59 years old
    - Each picture was labeled private or public if at least 75% of the judges were of the same opinion.
    - 4701 images are labeled as private; 27405 images are labeled as public; remainder are labeled as undecidable
- Features and Classifier
  - Features: SIFT
  - Classifier: SVM

User-defined Methods – Image Level

- An example of search results

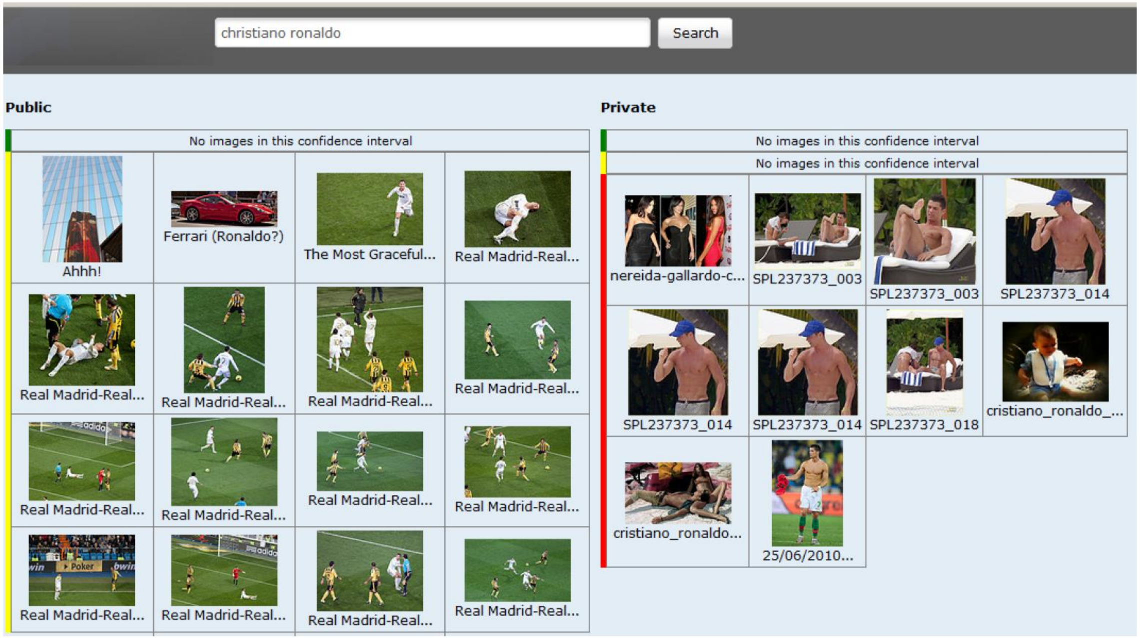


Figure 3: Private and public search results for the query "cristiano ronaldo" (June 06 2012).

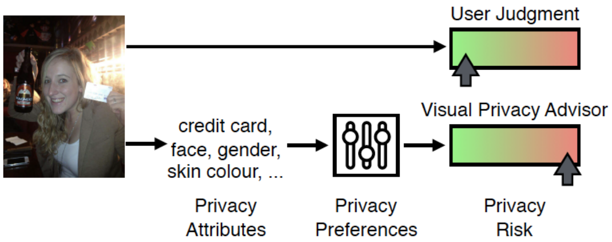
Privacy Detection Techniques

Take image, a representative modal of unstructured data, as an example:

- User-defined Methods
  - Image-level user survey
  - Object-level user survey
- Machine Learning Based Method
  - Content sensitiveness
  - Multimodal fusion

User-defined Methods – Object Level

- Towards a visual privacy advisor: Understanding and predicting privacy risks in images [2]
  - Aim to predict risk
  - Main work
    - Identify privacy attributes manually from multiple sources
      - EU Data Protection Directive 95/46/EC
      - US Privacy Act of 1974
      - Relevant attributes on Twitter/Flickr/Reddit/…
    - Label data manually (object-level)
    - Conduct user study to discover user privacy preference
    - Attribute detection & privacy risk prediction



User-defined Methods – Object Level

- The Visual Privacy(VISPR) Dataset: An object-level privacy dataset
  - Data Collection:
    - 22167 images from Flickr
  - Compilation of 68 privacy attributes from multiple sources:
    - EU Data Protection Directive 95/46/EC
    - US Privacy Act
    - The rules on prohibiting sharing personal information on various social networking websites (e.g., Twitter, Reddit, Flickr)
  - Data Labeling:
    - 68 privacy attributes of Images are labeled

User-defined Methods – Object Level

- Label distribution of VISPR Dataset

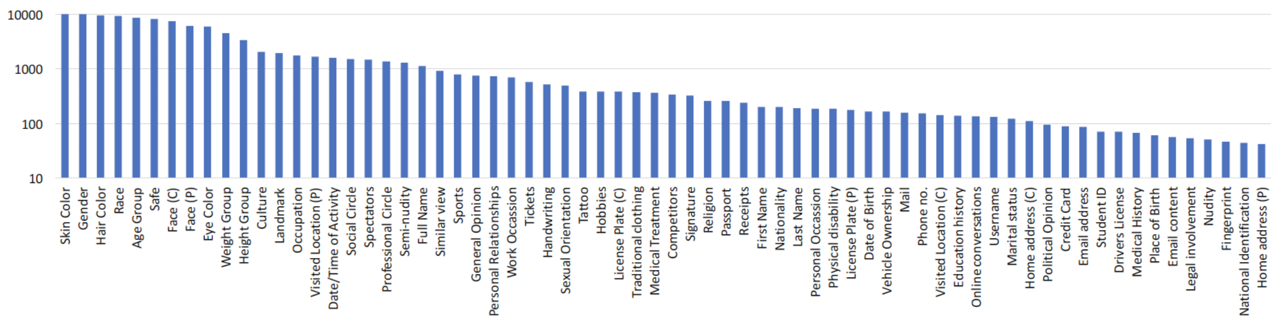
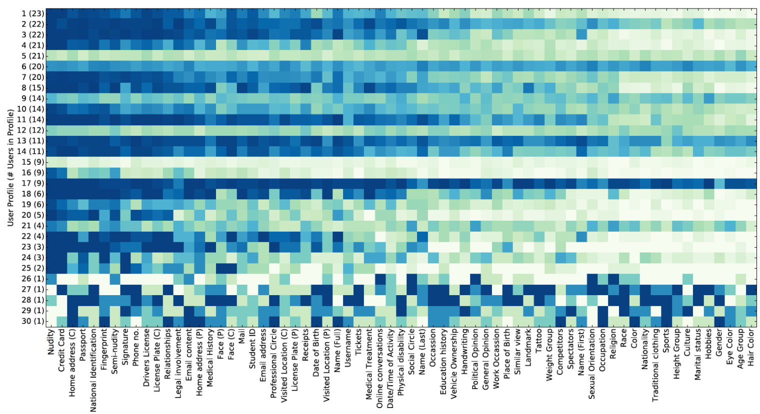


Figure 2: Label distribution in our dataset. Y-axis indicates the number of images.

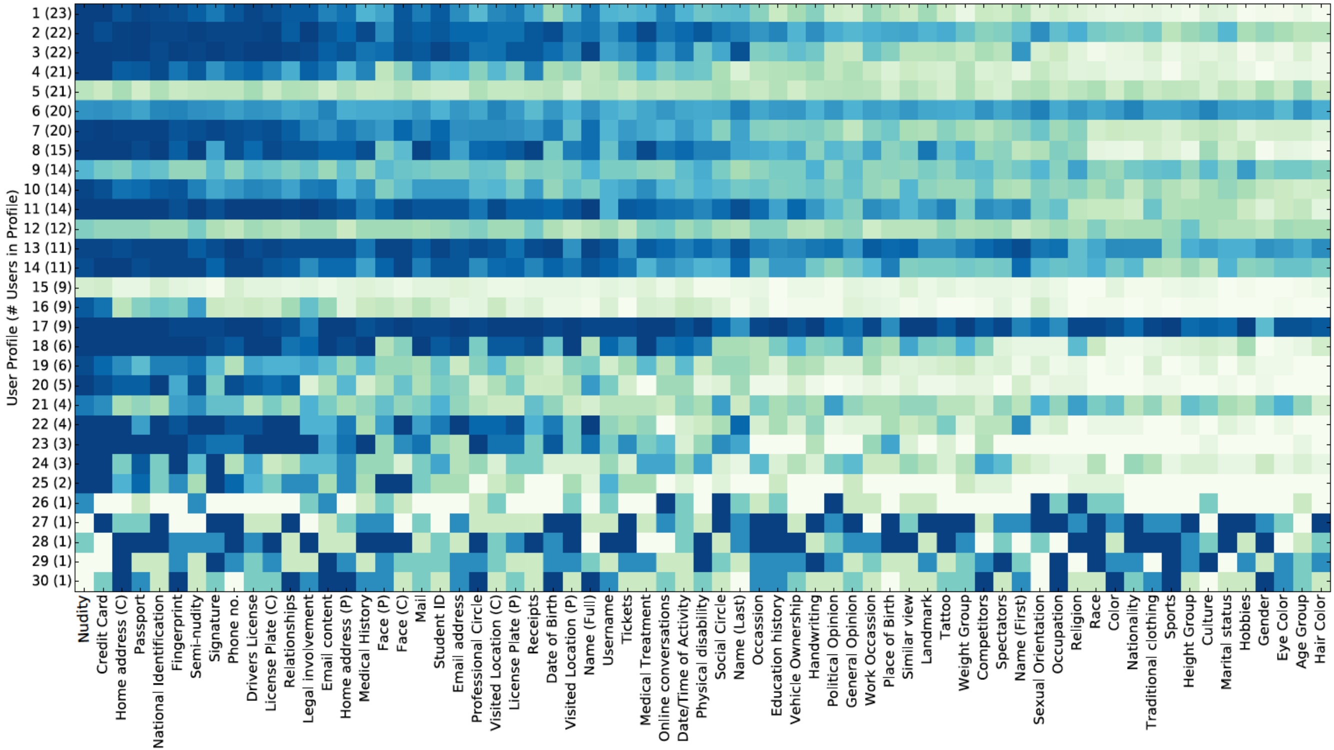
Images	22,167
Labels	115,742
Avg Labels/Image	5.22
Max Images/Label	10,460
Min Images/Label	44

User-defined Methods – Object Level

- User Study
  - 30 user profiles
    - 305 unique AMT workers
    - Workers rate images and 68 private attributes in VISPR independently
    - Higher score indicate greater sensitivity to privacy



User-defined Methods – Object Level



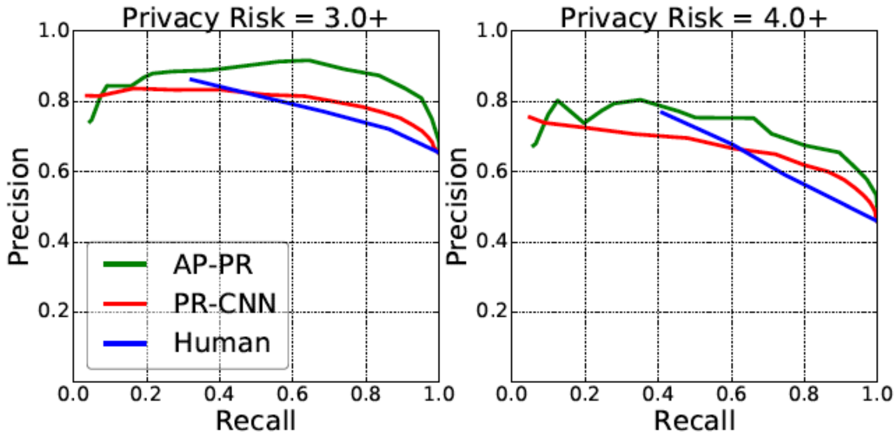
User-defined Methods – Object Level

- Privacy risk prediction
  - Identify user profile
  - Attribute detection
  - Privacy risk prediction

Definition 1. *Privacy Risk Score.* For some image $\boldsymbol{x}$, attributes $\boldsymbol{y} \in [0,1]^A$ and user preference $\boldsymbol{u} \in [0,5]^A$, the privacy risk score of image $\boldsymbol{x}$ containing attributes $\boldsymbol{y}$ on user $\boldsymbol{u}$ is $\max_a \boldsymbol{y}_a \boldsymbol{u}_a$

User-defined Methods – Object Level

- An interesting finding in VISPR
  - Users often fail to enforce their privacy preferences when sharing images online.



Privacy Detection Techniques

Take image, a representative modal of unstructured data, as an example:

- User-defined Methods
  - Image-level user survey
  - Object-level user survey
- Machine Learning Based Method
  - Content sensitiveness
  - Multimodal fusion

ML Based Methods – Content Sensitiveness

- iprivacy: Image privacy protection by identifying sensitive objects via deep multi-task learning[3]
  - Aim to predict privacy label of image
  - Main work:
    - Align privacy setting(Privacy/Public/Shared with acquaintance/...) to object classes
    - Detect private object classes via DNN
      - Use DNN to detect classes
      - Use top-down hierarchical clustering to improve efficiency

ML Based Methods – Content Sensitiveness

- Automatic object-privacy alignment

Feature Extraction →	Clustering →	Privacy alignment →	Refinement
Deep CNN	K-means	Frequency of occurrence	Random walk

ML Based Methods – Content Sensitiveness

Feature Extraction → Clustering → Privacy alignment → Refinement

Deep CNN | K-means | Frequency of occurrence | Random walk

- Extract feature by DNN
  - Full set of 1000 object classes
  - 1000-dimensional sparse representation vector *X*
- Clustering images according to KI
  - Number of clusters K is a parameter

$$\kappa_I(X_i, X_j) = \sum_{l=1}^{1000} \delta(X_i^l, X_j^l) \qquad \delta(X_i^l, X_j^l) = \begin{cases} 1, & if \quad X_i^l = X_j^l = 1; \\ 0, & otherwise \end{cases}$$

ML Based Methods – Content Sensitiveness

Feature Extraction	→	Clustering	→	Privacy alignment	→	Refinement
Deep CNN		K-means		Frequency of occurrence		Random walk

» Align privacy relevance score

- For each cluster:
  - Calculate relevance score between each object class and each privacy setting

$$\gamma(C_i, t) = \frac{\| \Psi(C_i, t) \|}{\| \Psi(C, P) \|} \quad t \in P$$

  - Align the privacy setting with highest γ to the object class

ML Based Methods – Content Sensitiveness

Feature Extraction	→	Clustering	→	Privacy alignment	→	Refinement
Deep CNN		K-means		Frequency of occurrence		Random walk

» Refine privacy relevance score

- Construct co-occurrence network
  - The object classes, which have large values of co-occurrences $\phi(.,.)$, are connected to form an object cooccurrence network.

$$\phi(C_i, C_j) = \rho(C_i, C_j) \log \frac{\rho(C_i, C_j)}{\rho(C_i) + \rho(C_j)}$$

$$\rho(C_i, C_j) = \frac{N(C_i, C_j)}{N}, \quad \rho(C_i) = \frac{N(C_i)}{N}, \quad \rho(C_j) = \frac{N(C_j)}{N}$$

ML Based Methods – Content Sensitiveness

- Refine privacy relevance score on co-occurrence network by random walk
  - Update score iteratively:

For each t, let $\rho_0(C_i, t) = \gamma(C_i, t)$

$$\rho_k(C_i, t) = \theta \sum_{C_j \in \Omega_{C_i}} \rho_{k-1}(C_i, t)\psi_{ij} + (1-\theta)\gamma(C_i, t)$$

where, $\psi_{ij} = \dfrac{\phi(C_i, C_j)}{\sum_{C_k \in \Omega_{C_i}} \phi(C_i, C_k)}$

- Align privacy setting t with biggest ρ to Ci
  - 268 object classes is identified as privacy-sensitive classes

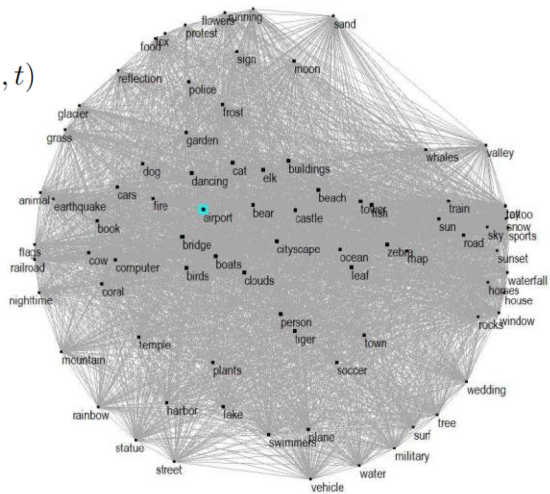


Fig. 6. **A small part of our object co-occurrence network.**

ML Based Methods – Content Sensitiveness

- Results of privacy alignment
  - Obtain 268 privacy-sensitive object classes from 1000 classes

TABLE I
The short list of privacy-sensitive object classes identif ed by this work.

Categories	Privacy-Sensitive Object Classes
Human Beings	portrait, people in birthday party, human body, human hair, human face, human eye, human neck, people in award, mannequin modeling, customer, $\cdots$
Family	baby, children, relatives/family, friend, husband, wife, parents, brother, sister, cousin, kids at play, african american, couple, $\cdots$
Woman	girl, explicit women, female surf ng, $\cdots$
Ethic	erotic, $\cdots$
House	home, bedroom, restroom, indoor, kitchen, $\cdots$
Clothes	suit, bikini, maillot, $\cdots$
Activity	drinking, wedding, swimming, bathing, working boys, sitting on boys, f shing, birthday parties, travel, vacations, fun summer vacation, $\cdots$
Work Lab	science lab, laptop, computer, personal , $\cdots$

ML Based Methods – Content Sensitiveness

- Results of privacy alignment
  - Obtain 268 privacy-sensitive object classes from 1000 classes

TABLE II

The short list of public object classes identif ed by this work.

Categories	Public Object Classes
Nature & Scenery	mountain, island, rock, sand, sea, coast, lake, river, sunset, sky, landscape, lakeside, sandbar, beach, cartoon, f re, ice, water, fashion, $\cdots$
Animal	pets, dog, cat, bird, wild animals, f sh, $\cdots$
Plant	f ower, tree, asian f oral, $\cdots$
Season	winter, sprint, summer, autumn, $\cdots$
Transportation	road, traff c, boat, car, $\cdots$
Building	House outside, garden, bridge,shopping center,park, bank, $\cdots$
Planet	moon, sun, earth, $\cdots$
City Signs	New York, Washington, Beijing, $\cdots$

Privacy Detection Techniques

Take image, a representative modal of unstructured data, as an example:

- User-defined Methods
  - Image-level user survey
  - Object-level user survey
- Machine Learning Based Method
  - Content sensitiveness
  - Multimodal fusion

ML Based Methods – Multimodal Fusion

» Dynamic Deep Multi-modal Fusion for Image Privacy Prediction[4]

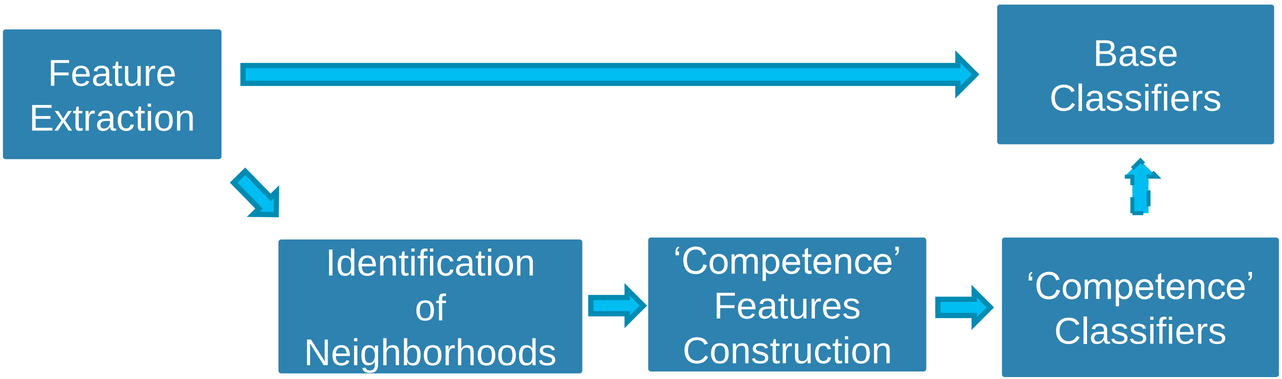
- Both image content, e.g. scene and object, and tags affect image privacy.
- Different images have different privacy factors

Single modality is correct			Multiple modalities is correct		
Image	Tags	Base classifiers	Image	Tags	Base classifiers
 (a)	bed, studio dining table speakers, music	**scene: 0.62** object: 0.5 tags: 0.29	 (d)	girl, baby indoor, people canon	scene: 0.49 **object: 0.87** **tags: 0.97**
 (b)	birthday night party, life	scene: 0.57 **object: 0.78** tags: 0.39	 (e)	people, party awesome, tea bed, blanket	**scene: 0.92** object: 0.38 **tags: 0.7**
 (c)	toeic, native speaker, text document, pen	scene: 0.02 object: 0.15 **tags: 0.86**	 (f)	indoor, fun party people	**scene: 0.92** **object: 0.73** **tags: 0.77**

ML Based Methods – Multimodal Fusion

» Dynamic fusion of three base classifiers

- Three base classifiers: Object, Scene and Tag classifiers
- Train ‘competence’ classifiers respectively to control the fusion process of base classifiers.



ML Based Methods – Multimodal Fusion

Feature Extraction → Identification of Neighborhoods → ‘Competence’ Features Construction → ‘Competence’ Classifier Training

Features

- Three base features by CNN: Train 3 base classifiers on the corresponding modality feature sets
  Object (F^o): 1000 dimension
  Scene (F^s): 365 dimension
  Image Tags (F^t): 265 dimension
- Combination feature: Using in ‘Neighoods identification’
  Object + Scene + Tag (F^{ost}): $f^{cat}(F^o, F^s, F^t)$
- Privacy profile feature: Using in ‘Neighoods identification’
  $$\overline{T} = \bigcup_{B_i \in \mathcal{B}} \{P(Y_T = private|T, B_i), P(Y_T = public|T, B_i)\}$$

ML Based Methods – Multimodal Fusion

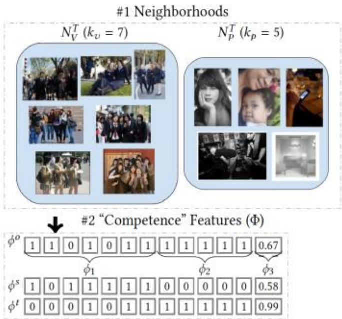
Feature Extraction → Identification of Neighborhoods → ‘Competence’ Features Construction → ‘Competence’ Classifier Training

Identification of Neighborhoods

- Find neighbors according to F^{ost} and $\overline{T}$ by cosine similarity.

‘Competence’ Features Construction

- $\phi^{o}, \phi^{s}, \phi^{t}$ are used to train corresponding competence classifiers
- Each ϕ is consist of three parts:
  - ϕ^{1} is the prediction results of Neighbors N_V^T by current base classifier
  - ϕ^{2} is the prediction results of Neighbors N_P^T by current base classifier
  - ϕ^{3} is the prediction results of input Image by current base classifier



ML Based Methods – Multimodal Fusion

Feature Extraction → Identification of Neighborhoods → ‘Competence’ Features Construction → ‘Competence’ Classifier Training

‘Competence’ Classifier Training

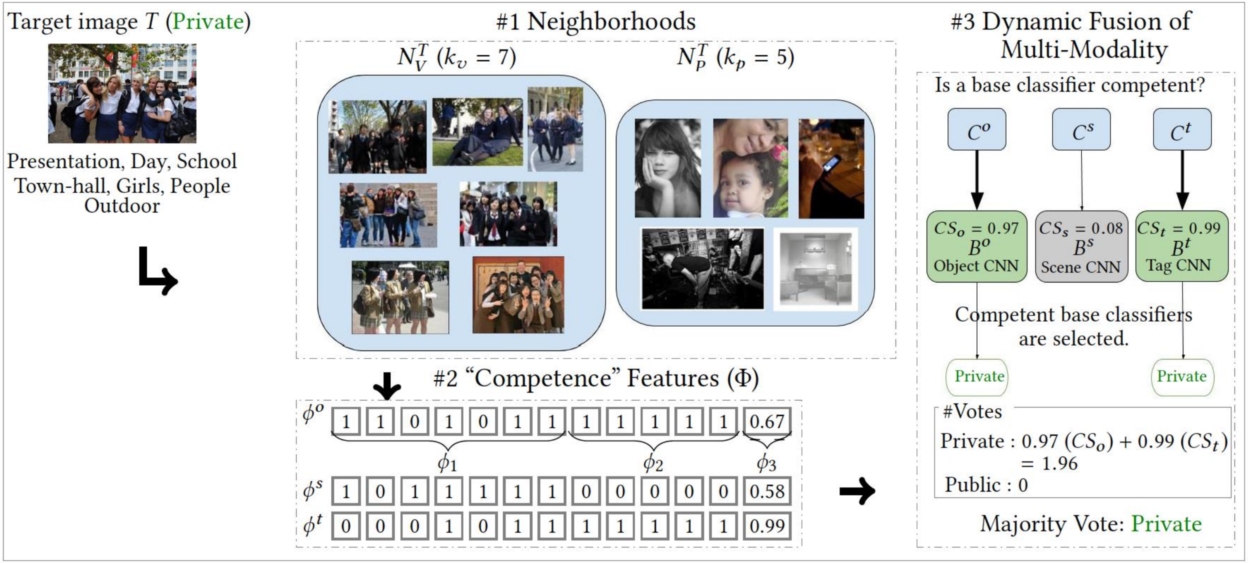
- Three competence classifiers $C = \{C^o, C^s, C^t\}$
- To train “competence” classifiers $C_i \in C$, we consider label Li = 1 if base classifier $B_i \in B$ predicts the correct privacy of a target image, otherwise 0.

Dynamic Fusion

- Dynamic voting: Dynamically determine the subset of most competent base classifiers
- Thresholding: If the competence score is greater than 0.5, then base classifier B_i is identified as competent to predict the privacy of target image
- The competence score is used as weight in final fusion

ML Based Methods – Multimodal Fusion

» An illustration of the proposed approach



ML Based Methods – Multimodal Fusion

Experiment Results

- Dataset is PicAlert

Features	Private			Public			Overall			
	Precision	Recall	F1-score	Precision	Recall	F1-score	Accuracy (%)	Precision	Recall	F1-score
DMFP	0.752	**0.627**	**0.684**	**0.891**	0.936	**0.913**	**86.36**	**0.856**	**0.859**	**0.856**
"Competence" Features										
DMFP−ϕ_1	0.777	0.553	0.646	0.874	0.951	0.911	85.74	0.849	0.852	0.844
DMFP−ϕ_2	0.74	0.565	0.641	0.875	0.939	0.906	85.11	0.842	0.846	0.84
DMFP−ϕ_3	0.752	0.627	0.683	0.891	0.936	0.913	86.35	0.856	0.859	0.856

Table 3: Evaluation of dynamic multi-modal fusion for privacy prediction (DMFP).

Model	 (a)	 (b)	 (c)	 (d)
DMFP	✓	✓	✓	✗
Object	✗	✓	✓	✗
Scene	✓	✗	✓	✗
Tags	✓	✓	✗	✗

Figure 4: Predictions for private images.

Reference

- [1] Zerr, Sergej, Stefan Siersdorfer, and Jonathon Hare. "PicAlert! a system for privacy-aware image classification and retrieval." Proceedings of the 21st ACM international conference on Information and knowledge management. 2012.
- [2] Orekondy, Tribhuvanesh, Bernt Schiele, and Mario Fritz. "Towards a visual privacy advisor: Understanding and predicting privacy risks in images." Proceedings of the IEEE International Conference on Computer Vision. 2017.
- [3]Yu, Jun, et al. "iPrivacy: image privacy protection by identifying sensitive objects via deep multi-task learning." IEEE Transactions on Information Forensics and Security 12.5 (2016): 1005-1016.
- [4] Tonge, Ashwini, and Cornelia Caragea. "Dynamic deep multi-modal fusion for image privacy prediction." The World Wide Web Conference. 2019.

Exercise

Exercise

- Ex.1
  - Read a paper about Privacy detection and write a report.
- Ex.2
  - Reproduce the deep random walk in [3].
- Ex.3
  - *Reproduce the dynamic fusion algorithm in [4]

THANKS!
Any questions?

You can find me at:

- @username
- user@mail.me