Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)

PrivacyAlert: A Dataset for Image Privacy Prediction

Chenye Zhao,[1] **Jasmine Mangat,**[2] **Sujay Koujalgi,**[3] **Anna Squicciarini,**[3] **Cornelia Caragea**[1]
[1] Department of Computer Science, University of Illinois at Chicago
[2] College of Information and Computer Sciences, University of Massachusetts Amherst
[3] College of Information Science and Technology, Pennsylvania State University
{czhao43, cornelia}@uic.edu, jmangat@umass.edu, {acs20,snk5290}@psu.edu

Abstract

Image privacy issues have become an important challenge as millions of images are being shared on social networking sites every day. Often due to users' lack of privacy awareness and social pressure, users' posted images reveal sensitive information and may be easily used to their detriment. To address these issues, several recent studies have proposed machine learning models to automatically identify whether an image contains private information. However, progress on this important task has been hampered by the absence of reliable, publicly available, up-to-date datasets. To this end, we introduce PrivacyAlert, a dataset developed from recent images extracted from Flickr and annotated with privacy labels (private or public). Our data collection process is based on state-of-the-art privacy taxonomy and captures a comprehensive set of image types of various sensitivity. We perform a comprehensive analysis of our dataset and report image privacy prediction results using classic and deep learning models to set the ground for future studies. Our dataset is publicly available at: https://doi.org/10.5281/zenodo.6406870.

Introduction

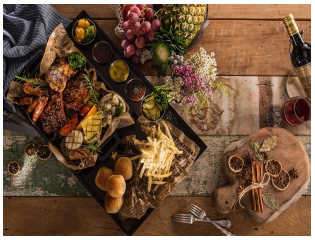
With the convenience brought by mobile Internet and mobile devices, the number of active users in online social networking sites is growing rapidly. People share personal aspects of their lives without much hesitation. With hundreds of millions of images being uploaded to various social networking sites every day (Yang et al. 2020), images have become one of the most prevalent forms among online social networking users (Tonge and Caragea 2019) for content sharing, but also a potentially risky one. For example, a seemingly common photo of a birthday party may unintentionally reveal sensitive information about a person's location, social relationships, and personal habits (Tonge and Caragea 2020). The address of an actress can be quickly located from images shared on her social networking site. This process can be done within 40 minutes with only Google Earth and basic geographical knowledge (Yang et al. 2020). In a TV show, the host learned the name and images of a little girl from the posts of her parents on their social networking sites. The host successfully identified the girl and gained her trust and took her away (Yang et al. 2020).

Although most social networking sites provide users with choices to set their privacy preferences, proper privacy setting is still a challenging and underperformed task for the vast majority of online users (Orekondy, Schiele, and Fritz 2017). Firstly, users find it troublesome to manually assign and manage privacy settings for each image they upload online (Zerr, Siersdorfer, and Hare 2012; Lipford, Besmer, and Watson 2008). Secondly, users often fail to follow through their own privacy preferences and disregard even simple privacy protection mechanisms (Orekondy, Schiele, and Fritz 2017). Therefore, it is unreliable and risky to solely rely on users for identifying sensitive content within their images.

The risk brought by online images has led to several studies proposing machine learning models for image privacy detection (Zerr et al. 2012; Squicciarini, Caragea, and Balakavi 2017; Tonge and Caragea 2020; Tran et al. 2016; Yang et al. 2020). In all these studies, images' content (and sometimes their tags) have been used as one of the discriminatory factors for learning images' sensitivity and related privacy designation. However, the absence of a publicly available, up-to-date dataset has limited the progress on this task. Zerr, Siersdorfer, and Hare (2012) developed a first gold-standard dataset in 2010, referred to as PicAlert. PicAlert includes images annotated with a private or public label by external annotators (private images are defined as ones that are on the private sphere and cannot be shared with everyone on the Internet. The rest are public). While pivotal to many of the recent studies in the space of online image privacy, PicAlert is no longer a representative repository for online image privacy research. In particular, images in PicAlert are at least 12 years old. The content that people are sharing online, peoples' awareness of privacy, and the underlying danger of privacy leakage has changed drastically over the years. In addition, data at the time was collected based on older studies on image privacy awareness, that dated even before the mass adoption of social networking and content sharing sites (Ahern et al. 2007). Other works investigating the problem of image privacy prediction (Yu et al. 2016; Tonge and Caragea 2019; Yang et al. 2020), have not yet provided publicly available datasets. For example, Yu et al. (2016) use CNNs to extract semantic image segmentation and learn the object-privacy relationship to identify privacy-sensitive objects. However, their dataset is not publicly available, nor is an in-depth discussion of the authors' approach

(a) children, face, joy

(b) Curtis, Passport



(c) food, kitchen table, eat, dishes

(d) Horse Teeth, Smile

Figure 1: Examples of private/public images from our dataset. (a) and (b) are private examples that reveal users' sensitive information. (c) and (d) are public examples that are safe to be shared online. We manually blur private images.

to data collection and labeling. Tonge and Caragea (2019) focus on developing a multi-modal privacy prediction model instead of proposing a new dataset. Orekondy, Schiele, and Fritz (2017) develop a dataset to predict privacy attributes that exist in each image, but it cannot be utilized to train binary privacy prediction models.

The aforementioned issues motivated us to develop a publicly available dataset that can represent the most up-to-date privacy patterns on social networking sites. We propose PrivacyAlert for image privacy prediction research. Although the definition of image privacy is subjective and often varies depending on peoples' personal traits and privacy awareness, some general privacy patterns that are commonly accepted exist, and are yet to be thoroughly understood through a current image dataset.

Our contributions can be summarized as follows: (1) We develop a publicly available labeled dataset for the task of image privacy prediction; (2) Our dataset includes a balanced selection of carefully selected images with sensitive content. Informed by recent work in the space (Orekondy, Schiele, and Fritz 2017; Li et al. 2018, 2020), we organize existing taxonomies of private images into 10 categories (e.g., nudity, personal information, facial expression, etc.). We present examples that are annotated as private and public in Figure 1. Figure 1a and 1b reveal a baby and a person's passport information, respectively, which are annotated as private. As examples of public content, we show photos of food (Figure 1c) and animals (Figure 1d); (3) We use the labeled dataset to obtain benchmark results using various types of deep learning models (CNN and transformer-based models). We perform experiments of not only image privacy prediction using single modalities (visual modality or textual modality), but also various types of multi-modal image privacy prediction (combination of visual and textual modalities); (4) We use Area Under the Margin (AUM) (Pleiss et al. 2020) based on deep learning models to automatically identify potentially mislabeled or ambiguous examples in the dataset. Interestingly, after removing these examples from the training set, we observe improvement in privacy prediction performance in deep learning models.

Related Work

Image privacy prediction is an important topic for online social networking sites (Zerr et al. 2012; Wu et al. 2018; Chandra et al. 2018; Kurtan and Yolum 2018; De Choudhury et al. 2009; Song et al. 2018). Ahern et al. (2007) study the problem of content and context-based image privacy. The authors claim that the research of image privacy prediction still requires more effort. This motivates the researchers to develop models that can automatically identify privacy patterns from images (Zerr et al. 2012; Squicciarini, Caragea, and Balakavi 2017; Zhong et al. 2017; Yu et al. 2016; Tran et al. 2016). Cruz et al. (2015) studies the correlation between image privacy and metadata (e.g., locations, time, shot details). Zerr, Siersdorfer, and Hare (2012) develop PicAlert dataset. The dataset consists of Flickr images posted from January to April in 2010. The images are then annotated by external users as private or public. However, the PicAlert dataset is no longer publicly available. And their random crawling data collection method cannot ensure a comprehensive dataset for private images. More recently, image privacy prediction is explored in many works (Yu et al. 2016; Yang et al. 2020; Tonge and Caragea 2019), but still a publicly available dataset that can represent general privacy patterns on up-to-date social networking sites is not available.

Zerr et al. (2012) use traditional visual features to train machine learning models to detect image privacy. Squicciarini, Caragea, and Balakavi (2017) utilize SIFT and tags to develop privacy classification models. With the development of deep learning, deep visual features can be extracted to reflect image privacy. Tonge and Caragea (2020) propose to use deep CNN networks pre-trained on ImageNet to extract visual features and use SVM as the privacy classifier. They also fine-tune the pre-trained CNN networks and yield better performance. Zhao and Caragea (2021) use BERT to learn deep features based on image tags and achieve a state-of-the-art performance of image tag-based image privacy prediction. Yang et al. (2020) propose to learn the object-privacy correlation using graph neural network for privacy prediction. As complementary information exists between images and tags, multi-modal privacy prediction networks are found to be able to further improve the performance of image privacy prediction. Tran et al. (2016) propose to use multiple visual modalities to predict image privacy. Tonge and Caragea (2019) assign competence scores for each modality to perform the weighted fusion. Our dataset can be used for the aforementioned complex explorations of deep learning models. We use our dataset to train and evaluate the aforementioned vision, language, and multi-modal models, then

compare the performance of these approaches.

Another line of research on image privacy prediction focuses on personalized privacy predictions which take the subjectivity in privacy definition into consideration. Spyromitros-Xioufis et al. (2016) propose a personalized image privacy prediction approach using deep neural networks. Zhong et al. (2017) argue that the bottleneck of reliable personalized image privacy prediction models is the limited user-specified data, and the time and space consuming to train and save models for each user. Moreover, users' deviations of sharing and privacy preferences can also influence the personalized model. Orekondy, Schiele, and Fritz (2017) define a set of privacy attributes as the taxonomy of private images based on social network rules. Then they use these attributes together with user preference to estimate privacy risk. However, Li et al. (2018) argue that the taxonomy based on public policies may not meet users' real needs. The authors propose another taxonomy of online private images based on online users (Li et al. 2020). In contrast to this direction, we focus on developing a dataset that can reveal the generalized privacy patterns. We utilize the privacy taxonomies (Orekondy, Schiele, and Fritz 2017; Li et al. 2020) in our data collection process.

Dataset

We collect our data using images with creative open common licence from Flickr, a popular social networking site for photo sharing. Flickr provides a public rich image repository. Each image is associated with tags (both user and system-generated) to describe and index the image. Flickr has been widely used by researchers for image data collection for various image processing tasks (Young et al. 2014; Lin et al. 2014). We use the API provided by Flickr that enables us to filter our crawling targets with dates and keywords. We then use the Amazon Mechanical Turk (AMT) crowd-sourcing platform for annotation.

Sampling Strategy

We collect images using an informed strategy. We begin with a targeted search addressing private images, per the taxonomy provided by (Orekondy, Schiele, and Fritz 2017). Here, authors identified 68 privacy *attributes* based on government policies and social networking sites rules on prohibiting sharing personal information. Privacy attributes indicate content that commonly exist in private images (e.g., credit card, race, etc.). Li et al. (2020) propose taxonomies of online private images as well as of sensitive content that may exist in each category. We combine the privacy attributes and the privacy categories from these two studies, and define a privacy taxonomy of *10* categories. Our categories are shown in Table 1. We use keywords for each category by way of image tags. For example, tags such as "bare", "body", and "naked", are most frequently used to describe images of the "nudity/sexual" category. We use these tags to define searching queries in Flickr API to crawl relevant images. Moreover, to develop a fresh and recent dataset, we focus on images posted in the recent 6 years: 83% of our dataset are images uploaded to Flickr from 2015 to 2021. The remaining 17% consists of images that are uploaded to Flickr from 2011 to 2015. Our crawling targets are restricted within the public domain. Specifically, we use the "Public Domain Dedication" and the "Public Domain Mark" licenses in Flickr API to do the crawling. We crawl 190,000 images and corresponding image tags (tags that attached by the users to describe image contents) using keywords of the 10 privacy categories. We then randomly select 20,000 images from the crawled images to perform annotation. We achieve a balanced number of samples for the 10 categories. This also enables us to analyze how image privacy is related to each category (i.e., what categories are more/less likely to be private).

Annotation

To annotate our data, we use the Amazon Mechanical Turk (AMT) crowd-sourcing platform. Each annotator is asked to read the following guideline upon accepting a task. The guideline is given as *"Assume you have taken these photos, and you are about to upload them on your favorite social network or content sharing site (e.g. Flickr, Facebook, Google+, Instagram). Please tell us whether these images are either private or public in nature. Assume that the people in the photos are those that you know"*. The annotators are then asked to classify each image into one of the four classes: clearly private, private, public, clearly public. Clearly private images are defined as images that should not be uploaded online at all. Private images are images that should be kept confidential for me and selected trusted people only. Public images are ones that anyone in my social network would be OK to see. Clearly public images are ones that anyone online would be OK to see. To complete the task, annotators are be able to see each image and assign one of the above labels.

To monitor the quality of annotation for each annotator, we create an attention checker set, as follows. We extract a sample of 50 private images and 50 public images annotated by 2 graduate students with the background of image privacy prediction. These images are selected from the most well-recognized types of private/public images (e.g., nudity for private, natural scenery for public). We randomly sample 1 private and 1 public image from the attention checker set and insert them into each annotation batch. If the annotator does not provide the expected answer on attention checkers, then the annotator's responses are dropped. 8% of annotators are dropped for failing to pass the attention checkers.

We first annotated 10,000 images using 3 annotators for each image as the training set. Then we annotate another 10,000 images as the validation/test set using 5 annotators for each image. The reason to use more annotators for validation and test is to increase annotations' reliability. Annotators are paid fairly for their work as per the standard local pay rate. Each annotator is paid 0.33 cents to label 22 images.

Inter-annotator Agreement

Traditional inter-annotator agreement measure such as Krippendorff's alpha would penalize fine-grained privacy classes equally, which is potentially problematic - as it would weigh

Category	Example	Keywords
Nudity/Sexual	People with full nudity or semi-nudity	bare, body, breasts, butt, erotic, naked, nudity, sexual, shirtless, kissing
Other people	Photos of or with people one knows/ the owner/bystanders/events	grandparent, children, spectator, boy, family, husband kids, parents, partner, people, wife, family
Unorganized home	Messy/unorganized home	messy room, toilet, uncleaned pool, bathroom, disorganized, restroom, unclean, indoor, bedroom, kitchen, living room, desk, sofa,trash, closet
Violence	Photos of violent scene	damage, guns, violence, war, military, shooting, firearms weapons, corps, battlefield, tourisms, theater, legal, license
Medical	Photos of medical condition/visible blood/medical treatment	eye, abscess, peeling skin, bad teeth, acne, bloody injury, wound, surgery, peel, pharmacy, emergency, tongue, gummy, throat, lip, infection, pain
Drinking/Party	Photos of a party/drinking/smoking	drinking, body shot, drunk, hang out, smoking, alcohol party, cigarette, music, event, concert, night, vodka
Appearance/Facial expression	Photos of appearance/facial expression/ pose/clothing that reflect negatively on personal character	tattoo, ungroomed, messy hair, overweight,piercing, unflattering appearance, unsatisfying body, unfashionable funny looking, strange hair, wig, scary looking, silly forced smile, unamused face, tight clothing
Bad character/ Unlawful criminal	Crime scene/unlawful behaviours	infidelity, cheating, dangerous, illegal, drugs, abused arrest, children in danger, mugshot, marijuana, drug, kills, thief, stealing, bomber, smuggler, gang, prisons, robbery
Religion/Culture	Photos that reveal people's religion belief	culture, religion, spiritual, bible, catholic christian, christianity, church, faith, hinduism, holy indian, islam, judaism, religious, sacred, sikhism, traditional
Personal information	Personally identifiable information	bank account, home address, license plate, automobile, credit card, email, passport, password, sign, ticket, laptop browser, computer, internet, railway, flight, username

Table 1: Privacy taxonomy and keywords table.

the distance between a "private" and "public" annotation and a "clearly private" and "private" annotation equally. We expect an image with a private and a public annotation to have larger disagreement scores compared with a private and a clearly private annotation. Inspired from (Desai, Caragea, and Li 2020), we incorporate the distance between the fine-grained privacy annotations into the agreement score and define the pairwise agreement (PA) between workers as follows:

$$PA_{avg} = \frac{1}{N}\sum_{n=1}^{N} PA_n \tag{1}$$

$$PA_n = \frac{2}{m(m-1)}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} d(w_i, w_j) \tag{2}$$

$$d(w_i, w_j) = 1 - |f(e_i) - f(e_j)| \tag{3}$$

where N is the total number of images. PA_n is the average of all pair-wise distances. m is the number of annotators. f maps the worker w_i's annotation (e_i) and w_j's annotation (e_j) to numbers: clearly private: 0, private: 0.25, public: 0.75, clearly public: 1.00. This mapping reflects the difference between annotations: private is "more different" from public ($|0.25 - 0.75|$) than clearly private ($|0.25 - 0|$). Distance is inversely proportional to the agreement score (as shown in (3)). The correctness of this pairwise agreement metric is verified from a user study performed in a similar approach (Desai, Caragea, and Li 2020). Here, we observe a similar PA score obtained using 3 and 5 annotators. When using 3 annotators, we obtain an average value of 0.81 on all classes, which shows high agreement.[1] We also study the per-class inter-agreement score, and observe lower inter-annotator agreement on the private-sphere classes (clearly private: 0.66, private: 0.56) than the public-sphere classes (public:0.80, clearly public: 0.89). This result implies that annotators agree more on the public images, but tend to be more subjective when annotating private images. Among 10 privacy categories, we observe that "Nudity/Sexual" and "Appearance/Facial Expression" show high PA scores for the private classes: 0.82 and 0.71, 0.71 and 0.74 for clearly private and private, respectively. In contrast, "Bad Characteristics/Unlawful/Criminal", "Personal Info", and "Religion/Culture" show high agreement scores for the public classes: 0.81 and 0.91, 0.80 and 0.92, and 0.81 and 0.91 for public and clearly public, respectively. We also transform the 4-class annotations into binary labels and compute the inter-annotator agreement using Fleiss' kappa, and obtain an average value of 0.37 on all privacy classes.

Analysis

Privacy Distribution We explore both the overall privacy distribution and category-wise privacy distribution in order to gain a better understanding of which specific categories are identified and perceived as private. Results are shown in Figure 2.

From Figure 2, we observe that public and clearly public

[1] A reasonable interpretation of PA scores may be as follows: 0—0.25 (no agreement), 0.25—0.50 (poor agreement), 0.50—0.75 (moderate agreement), 0.75—1.00 (high agreement).

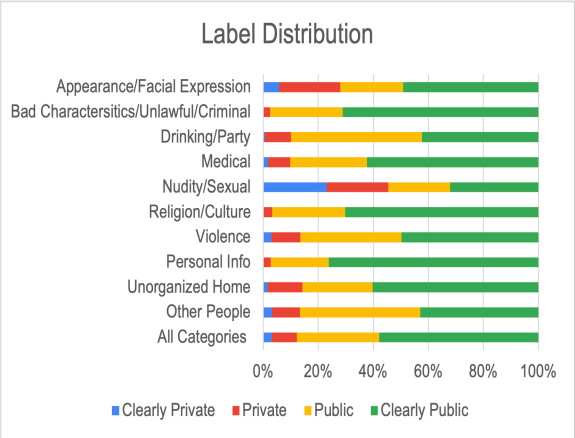


Figure 2: Label distribution.

categories take the majority of all annotated images. Overall, 3% and 9% are annotated as clearly private and private, respectively. 30.3% and 57.6% of images are annotated as public and clearly public. The prevalence of public images is not surprising, as our data collection targets publicly available images. With respect to category-wise distribution, we observe that "Nudity/Sexual", "Appearance/Facial Expression" show a higher percentage of images annotated with a private label (including clearly private and private). Particularly for the "Nudity/Sexual" category, the breakdown is: 23.00%, 22.36%, 22.49%, and 32.15%, for clearly private, private, public, and clearly public, respectively. In contrast, images belonging to the "Bad Characteristics/Unlawful/Criminal", "Personal Info", and the "Religion/Culture" categories show a very low percentage of images on the private side (less than 4%).

(a) Private

(b) Public

Figure 3: Examples of images with same tags ("redeployment", "welcome home", "ceremony") and annotated with different privacy labels

Annotators' Consistency Images uploaded with a same set of tags (e.g. in an album) are typically semantically related (e.g. a party, an event, a trip). Therefore, we expect these to be annotated with a same privacy rating. However, in our dataset we identify many cases where the images labeled under the same tags have been annotated differently. Our dataset has 4000 images that share the same tags with other images. Among them there are 1853 unique tags. Out of those, 19.10% (354) unique tags have images with different annotations. An example can be found in Figure 3, where Figure 3a and 3b are uploaded together and share same image tags but they are annotated with different privacy annotations.

	Train	Validation	Test
Private	788	466	450
Public	2348	1398	1350
Overall	3136	1864	1800

Table 2: Train, validation, and test splits for PrivacyAlert.

To better understand the reasons for this inconsistency, we consider two potential explanations: either the annotators are inattentive or inaccurate, or the images, despite their labels, are not visually or contextually similar. To confirm the first hypothesis (i.e. annotators' quality is low), we first checked the annotation tasks. Images were randomly assigned to the workers. Every image grouped under the same tag was annotated by different workers. This means that each worker can have their privacy preference of an image. Moreover, of the workers who had annotated multiple images under same tags, 65.22% of them were consistent with their annotations. For example, if a worker had annotated 2 or more images with same tags, then 65.22% of the time he had annotated the images as either private or public. To confirm the second hypothesis, we first use the output of the last fully-connected layer of ResNet101 pre-trained on ImageNet to extract visual features for each image. We then calculate the 10 closest neighbors for each image using the cosine similarities of extracted features. We found that there are only 99/354 unique tags with images visually similar to each other. This result suggests that sharing tags does not imply visual similarity - and therefore explains the privacy annotation disagreement.

We also investigate whether visually similar images tend to have similar privacy annotations. For each image, we extract visual features using pre-trained ResNet and calculate its 5 closest neighbors. We observe that 62.5% of the images have more than 3 out of 5 visually similar images annotated with the same privacy label.

Benchmark Dataset We focus on the task of binary image privacy prediction, consistent with prior works. Accordingly, we transform the 4-class annotations into binary labels.[2] We first integrate votes to the private, public sphere, respectively. i.e., we combine votes for *clearly private* and *private* as votes for the *private* class, and combine votes for *public* and *clearly public* as votes for the *public* class. The final annotation of the image is the class (private or public) with the highest votes. We first split our dataset into training, validation, and test sets with the size of 10000, 5000, 5000, respectively. We remove samples with the same image tags as they cannot be used to train image tag-based models. Public images take the majority part of the online images (the number of public and private images are in the ratio of 9:1). To balance the dataset, we downsample the public set. The balanced ratio of public and private is 3:1. The final dataset split is shown in Table 2.

Baseline Modeling

We model PrivacyAlert dataset using the following schemes: (1) single-modal privacy prediction, where we use one

[2]Note that our PrivacyAlert dataset offers fine-grained privacy classification capability that we plan to explore in the future.

	Model	Private			Public			Overall			
		Prec	Rec	F1	Prec	Rec	F1	Acc %	Prec	Rec	F1
Object	ResNet101+SVM	0.583	**0.680**	0.628	0.887	0.838	0.862	79.83	0.811	0.798	0.803
	ResNet101+FT	**0.726**	0.667	**0.694**	**0.892**	**0.916**	**0.904**	**85.39**	**0.850**	**0.854**	**0.852**
Scene	ResNet50+SVM	0.639	**0.644**	0.642	**0.881**	0.879	0.880	82.00	0.821	0.820	0.820
	ResNet50+FT	**0.699**	0.613	**0.653**	0.876	**0.912**	**0.894**	**83.70**	**0.832**	**0.837**	**0.834**
Tags	TagCNN (UT)	0.671	0.703	0.688	0.900	0.884	0.892	83.94	0.843	0.839	0.841
	BERT+SVM (UT)	0.569	0.649	0.606	0.877	0.836	0.856	78.94	0.800	0.789	0.794
	BERT+FT (UT)	0.700	0.684	0.692	0.896	0.902	0.899	84.78	0.847	0.848	0.847
	BERT+FT (DT)	0.728	0.638	0.680	0.884	**0.921**	0.902	85.00	0.845	0.850	0.847
	BERT+FT (UT+DT)	**0.746**	**0.704**	**0.725**	**0.903**	0.920	**0.912**	**86.61**	**0.864**	**0.866**	**0.865**

Table 3: Comparisons of single-modal image privacy prediction models. FT means fine-tuning. UT, DT, and UT+DT represent user tags, deep tags, and the combination of user tags and deep tags, respectively.

modality (visual-only or textual-only) to predict image privacy; (2) multi-modal privacy prediction, where privacy prediction is made based on multiple modalities (visual and textual modalities); (3) Area Under the Margin (AUM) with deep learning models to identify mislabeled samples from our dataset.

Single-modal Models

Object-based single-modal privacy prediction models. Object information is found to be a fundamental modality to access the privacy nature of an image. For example, a single element such as underwear, license plate, and credit card can be a strong indicator of a private image. Recently, pre-trained deep learning models have risen in popularity. The knowledge learned from the pre-trained task can be utilized for downstream tasks. In our work, we adopt ResNet101 pre-trained on ImageNet for object identification into the task of image privacy prediction. We experiment with two approaches: (1) pre-trained model is adopted only to extract features from input images. The extracted features are then used to train an SVM for image privacy classification; (2) pre-trained model is fine-tuned for privacy prediction: we change the output units from 1000 (object categories in ImageNet) to 2 (binary privacy classes). We initialize the weights of the modified model with its pre-trained counterpart and fine-tune the model using PrivacyAlert for privacy prediction.

Scene-based single-modal privacy prediction models. As consistently shown in previous works (Ahern et al. 2007; Tonge, Caragea, and Squicciarini 2018), the scene context of an online image can also imply privacy. For example, images about public places and home environments may be linked with different privacy preferences. We adopt ResNet50 pre-trained on Place365 dataset (Zhou et al. 2017) which contains images annotated as 365 scene classes (e.g., dressing room, airport, etc). Similarly, we experiment with (1) using the pre-trained ResNet50 to extract scene descriptors of input images from the last fully-connected layer and train an SVM for privacy classification; (2) replacing the unit of the last fully connected layer from 365 (scene categories) to 2 (privacy categories). Then we fine-tune the model.

Image tag-based single-modal privacy prediction models. Image tags are words attached by online users to describe the image content, which is found to be another good indicator of image privacy (Zerr et al. 2012; Squicciarini, Caragea, and Balakavi 2017). For example, an image with tags "bra, dressing room" and another image with tags "mountain, nature" may have different privacy orientations. Previous works use CNN for image tag-based privacy classification (Tonge and Caragea 2020). Recently, Zhao and Caragea (2021) propose to use BERT for this task and achieve state-of-the-art performance. In our work, we experiment with CNN and BERT for image tag-based privacy prediction. We use user tags which are user-attached tags that are directly crawled from the web. As the set of user tags may be very sparse and often contain noisy words (Sundaram et al. 2012), we extract deep tags (Tonge and Caragea 2020) from images. We use the top 10 object categories as deep tags from the probability distribution extracted from pre-trained CNN.

Multi-modal Models

PCNH. PCNH (Tran et al. 2016) consists of two joint architectures: an AlexNet which extracts object features from input images, and a CNN model which extracts convolutional features. The two types of features are combined for privacy classification.

Concat. This approach (Tonge, Caragea, and Squicciarini 2018) uses AlexNet models pre-trained on ImageNet and Place365 to extract object and scene tags, respectively. Particularly, top k object category names with the highest probabilities for each input image are selected as object and scene tags. The combination of object tags, scene tags, and user tags is used to train an SVM classifier for image privacy prediction.

DMFP. DMFP (Tonge and Caragea 2019) fuses single-modal predictions using a weighted majority vote. Fusion is weighed by competence scores for every single modality, which is generated by competence classifiers learned from neighborhood information of each input example. The predictions made by basic classifiers are fused for privacy classification.

VilBERT. VilBERT (Lu et al. 2019) is a pre-trained model for vision-language tasks. VilBERT is based on the transformer model, that shares the same two-stream BERT with

	Private			Public			Overall			
	Prec	Rec	F1	Prec	Rec	F1	Acc %	Prec	Rec	F1
PCNH	0.706	0.511	0.593	0.851	0.929	0.888	83.17	0.831	0.832	0.831
Concat	0.626	0.716	0.668	0.900	0.858	0.879	82.22	0.832	0.822	0.826
DMFP	0.666	0.656	0.661	0.886	0.890	0.888	83.17	0.831	0.832	0.831
VilBERT	0.658	0.697	0.677	0.897	0.879	0.888	83.37	0.837	0.834	0.835
Gated	**0.779**	**0.722**	**0.750**	**0.910**	**0.932**	**0.921**	**87.94**	**0.877**	**0.879**	**0.878**

Table 4: Multi-modal image privacy prediction results.

	Private			Public			Overall			
	Prec	Rec	F1	Prec	Rec	F1	Acc %	Prec	Rec	F1
Gated	0.565	0.622	0.592	0.870	0.840	0.855	78.56	0.793	0.786	0.789
Object	0.510	0.707	0.592	0.888	0.773	0.827	75.67	0.793	0.757	0.768
Scene	0.535	0.624	0.576	0.867	0.819	0.843	77.06	0.784	0.771	0.776
Tag	0.526	0.668	0.587	0.879	0.800	0.837	76.68	0.791	0.767	0.775

Table 5: Prediction performance of models that are trained on PicAlert and test on the test set of PrivacyAlert.

a co-attention scheme to fuse the visual and textual information. VilBERT has achieved state-of-the-art performance on many vision-language tasks. We fine-tune VilBERT using images and tags as a strong baseline for image privacy prediction.

Gated Fusion. Gated fusion is a decision-level multi-modal fusion technique. Privacy predictions generated by the trained single-modal models are fed into the Gated fusion network. Gated network dynamically learns fusion weights from single-modal predictions representing the reliability of each single modalities. The fusion weights are then used to regularize the weighted average fusion so that more reliable modalities are strengthened with higher weights, and less reliable modalities are restrained with lower weights.

AUM

In many real-world scenarios, datasets may contain samples that are mislabeled. Even in the most celebrated datasets such as MNIST and ImageNet, some harmful examples also exist. This is because human annotators are prone to make mistakes. Mislabeled training data introduces noise to deep neural networks: the model may achieve zero training error for those mislabeled data or generate random predictions. In our case, a private image that is mislabeled as public may confuse the model and limit the performance for the private image prediction, which further leads to privacy leakage. To tackle this issue, we use AUM (Area Under the Margin) (Pleiss et al. 2020) based on deep learning models to identify mislabeled examples from the training set. Specifically, when an input example is fed into a deep learning model, AUM measures the average distance between the logits values for the example's predicted class and its non-assigned class.

AUM for a training sample $\mathbf{x}$ is calculated as:

$$AUM(\mathbf{x}, y) = \frac{1}{T}\sum_{t=1}^{T}(z_y^{(t)}(\mathbf{x}) - z_i^{(t)}(\mathbf{x})) \quad (4)$$

Category	F1-overall	F1-private	F1-public
Overall	0.878	0.750	0.921
Nudity/Sexual	0.813	**0.853**	0.752
Other People	0.827	0.683	0.892
Unorganized Home	0.950	0.640	0.974
Violence	0.854	0.623	0.913
Medical Cond/Blood	0.909	0.727	0.945
Drinking/Party	0.846	0.468	0.921
Appear/Facial Express	0.842	0.757	0.883
Bad Character/criminal	**0.972**	0.500	**0.989**
Religion/Culture	0.950	0.444	0.977
Personal Info	0.950	0.250	0.974

Table 6: Per-category privacy prediction performance using the Gated fusion model.

where y is the class label. T is the total number of training epochs. $z_y^{(t)}(\mathbf{x})$ and $z_i^{(t)}(\mathbf{x})$ represent the logits of the predicted class and the non-predicted class, respectively. For correctly labeled examples, which are generalized from similarly-labeled data, the difference between the two logits should be large, because the class label can be clearly distinguished by the model. Examples with low difference between the predicted class and the non-predicted class could be examples that are difficult to be identified by the deep learning model or ones that are mislabelled. To split the "mislabeled" and the "difficult" examples, we randomly select a subset of the training set as the threshold set (Pleiss et al. 2020) and assign them an extra non-exist class label and train the model. Samples that have lower AUM scores than the majority of the threshold set are mislabeled data and are removed from the training set.

Experiments and Results

In this section, we describe the details of our image privacy prediction experiments and results.

Experimental Settings

We report both the overall and class-wise (private and public) performance. The metrics we adopt include overall pre-

		Private			Public			Overall			
	Mode	Prec	Rec	F1	Prec	Rec	F1	Acc %	Prec	Rec	F1
Gated	Ori	0.779	**0.722**	0.749	0.910	0.932	0.921	87.94	0.877	0.879	0.878
	AUM	**0.796**	0.711	**0.751**	**0.907**	**0.939**	**0.923**	**88.22**	**0.879**	**0.882**	**0.880**
Object	Ori	0.726	0.667	0.694	0.892	0.916	0.904	85.39	0.850	0.854	0.852
	AUM	**0.744**	**0.711**	**0.727**	**0.905**	**0.919**	**0.912**	**86.67**	**0.865**	**0.867**	**0.866**
Scene	Ori	**0.699**	0.613	0.653	0.876	**0.912**	**0.894**	83.70	0.832	0.837	0.834
	AUM	0.682	**0.662**	**0.672**	**0.888**	0.897	0.893	**83.80**	**0.837**	**0.838**	**0.838**
Tag	Ori	0.746	0.704	0.725	0.903	0.920	0.912	86.61	0.864	0.866	0.865
	AUM	**0.759**	**0.716**	**0.737**	**0.907**	**0.924**	**0.916**	**87.22**	**0.870**	**0.872**	**0.871**

Table 7: Comparisons of single-modal classifiers (i.e., fine-tune ResNet101, ResNet50, and BERT) and Gated fusion model with AUM. Ori means AUM is not applied.

diction accuracy, and F1-score, precision, and recall for overall, private class, and public class, respectively. All hyper-parameters are selected on the validation set. We use Linear SVM with squared hinge loss.

Results

Single-modal Approaches. We experiment with single-modal image privacy prediction models: object, scene, and image tags. The results are shown in Table 3. We observe following patterns. Firstly, directly fine-tuning (FT) deep learning models consistently show better performance for single-modal image privacy prediction than SVM classifiers. Secondly, the combination of user tags and deep tags (UT+DT) outperform either of them individually. We use the combination of user tags and deep tags based on BERT for the rest experiments. Thirdly, the best performing tag-based classifier (BERT+FT(UT+DT)) shows the best performance among three modalities, illustrating that in our dataset, image tag is the most effective type of information source that reveals image privacy.

Multi-modal Approaches. Results of multi-modal image privacy prediction approaches are shown in Table 4, where we observe that the Gated fusion model consistently achieves the best performance on all compared metrics. The improvement of Gated over PCNH shows the importance of scene and image tags for multi-modal privacy prediction. The improvement of Gated over Concat illustrates the advantage of weighted average fusion. The improvements from Gated to DMFP illustrates that fusion weight can be learned in a more efficient yet effective way from single-modal predictions. Gated also outperforms VilBERT. Our inspection discloses that VilBERT needs more data to learn the image-tag correlation.

Privacy prediction models trained on PicAlert and evaluated on PrivacyAlert. PicAlert captures images posted over 12 years ago, which may not represent the image privacy patterns of the current social networking sites. To investigate this problem, we train deep learning models using PicAlert dataset and evaluate them using test images of PrivacyAlert dataset. We experiment with our best-performing single-modal and multi-modal approaches: fine-tuning ResNet101, ResNet50, and BERT as single-modalities for the object, scene, and image tags, respectively, as well as the Gated fusion network for multi-modal fusion. Results are shown in Table 5. We observe that when trained with PicAlert and tested with PrivacyAlert, the performance of all models drop significantly compared with models that are trained with PrivacyAlert. This result suggest that PicAlert cannot reflect privacy patterns of current online social networking sites.

Per-category Performance We calculate the privacy prediction performance of the Gated fusion model for each privacy category in the test set. The results are shown in Table 6. From the results, we observe the following patterns. Firstly, "Nudity/Sexual" and "Appearance/Facial Expressions" show high F1-private. These two categories also have the highest percentage of private examples. Comparatively, "Personal Information", "Drinking/Party", "Religion/Culture" and the "Bad Characteristics/Unlawful/Criminal" show the lowest F1-private, which are the categories that have the lowest percentage of private examples. This suggests that the privacy categories with high percentage of private examples can help the privacy prediction model to learn better patterns of private images and make better predictions, whereas the privacy prediction model can not predict well for private class on categories with low percentage of private examples, although they may achieve good performance on the public class.

AUM In our experiment, we randomly select and assign a subset ($\frac{1}{3}$ of the training set) with an artificial privacy class that does not exist. This subset is referred to as the threshold sample set (Pleiss et al. 2020) and used to mimic the mislabeled data. The 99^{th} percentile of AUM scores of the threshold sample set are treated as the threshold between correctly and mislabeled data (Pleiss et al. 2020). Training data that yields lower AUM than this threshold are mislabeled samples. We identify around 100 mislabeled training samples. We then remove those mislabeled samples from the training set. Then we use the updated training set to train single-modal classifiers. Experimental results are shown in Table 7. We observe that, after identified mislabeled samples are removed from the training set, all single-modal classifies achieve better performance, especially in the private class (e.g., F1). We then use the updated single-modal predictions to train the Gated fusion network. The results show that improved single-modalities further contribute to the improvements on the multi-modal network: Gated with AUM on single-modalities improves its original counterpart.

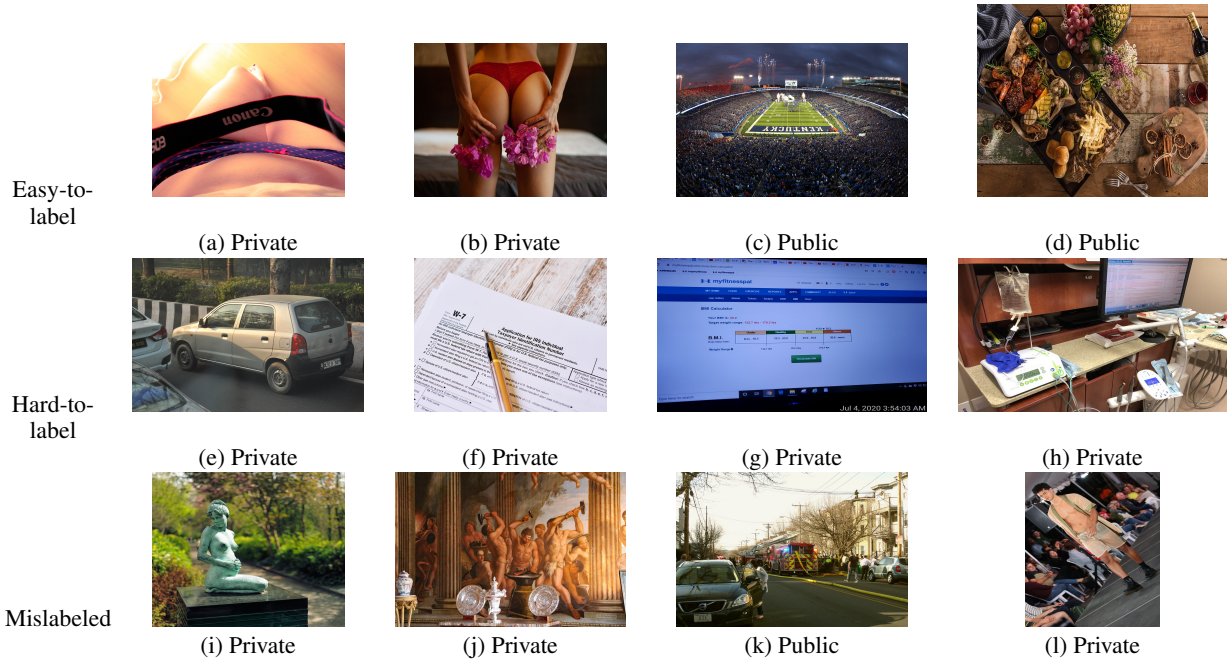


Table 8: Examples of easy-to-label, hard-to-label, and mislabeled training samples identified by AUM based on ResNet101.

Examples Detected by AUM

To better understand the effectiveness of AUM, we analyze training samples that are assigned with high AUM scores (easy-to-label). These refer to training samples that are easy to be identified by deep learning models. We also study training samples with low AUM scores which can be either hard-to-label or mislabeled samples. Hard-to-label samples are ones that are annotated correctly, but their privacy pattern are hard to be identified by deep learning models. Mislabeled samples are ones that have wrong annotations caused by inaccurate annotators. We gather samples for each type in Table 8. We observe that for easy-to-label private samples ((a) and (b)), private information is mostly revealed from the visual information (e.g., underwear), which can be easily identified by ResNet101. Public images with high AUM scores ((d) and (e)) are mostly about public places or objects (i.e., stadium and food). Hard-to-label examples are shown in the image (e)-(h). We observe that the privacy information of this type of images is not directly reflected by visual features. Some require additional background knowledge or common sense. For example, the license plate number of the car (image (e)), the tax document (image (f)), the body height and weight information (image (g)), as well as the patients' information on the monitor (image (h)) make the images to be private. Such privacy information is hard to be captured by ResNet101. As for mislabeled images, we observe that images of sculptures and paintings with body exposure that are visually similar to real human beings are often mislabeled as private images (image (i) and (j)). Moreover, images whose private information is visually hard to be observed by the annotators are often mislabeled. For example, the license plate number of the car on the corner (image (k)). Another type of mislabeled image is the image whose privacy is decided by multiple types of modalities. For example, image (l) is a model on the stage. Although the object itself may be considered as private, we can still decide the image as public by considering the scene context of the image (a public activity). AUM identifies such challenging image and helps us better utilize the training samples.

Conclusion

Privacy leakage from images that people shared on social networking sites has become an important challenge, acknowledged by both researchers and practitioners. Consumers' lack of privacy awareness and of privacy mechanisms to support their preferences has led to several machine learning and deep learning tools to learn privacy patterns from user-owned images. However, research progress in this space has been hampered by the absence of a publicly available, up-to-date image dataset. Our study focuses on creating a dataset that is representative of recent privacy patterns on social networking sites, and can be used to train image privacy classifiers based on various types of machine learning and deep learning models. We use state-of-the-art privacy taxonomies as the guideline for our data collection process. As a result, our dataset includes several types of potentially private images. We report classification benchmark results on various types of deep learning models trained with single-modal and multi-modal strategies. Our results can be compared in future studies. We also use AUM to identify mislabeled data to improve the quality of the training set. To this end, PrivacyAlert is the most up-to-date dataset for the task of general image privacy prediction for the research community. Our dataset also has the potential to be utilized for domain adaptation between privacy categories.

Ethical Statement

We use the "Public Domain Dedication" and the "Public Domain Mark" licenses in Flickr API for image crawling. Accordingly, our dataset only includes images with a public license. Also, our dataset only includes images and tags, every image is privy to potentially personally identifying metadata. Images are collected using generic keywords instead of user information as queries, therefore our dataset does not have a large collection of images from an individual user.

Acknowledgements

This work was partially supported by the National Science Foundation. The computation for this project was performed on Amazon Web Services through a research grant to UIC.

References

Ahern, S.; Eckles, D.; Good, N. S.; King, S.; Naaman, M.; and Nair, R. 2007. Over-exposed? Privacy patterns and considerations in online and mobile photo sharing. In *SIGCHI 2007*, 357–366.

Chandra, D. K.; Chowgule, W.; Fu, Y.; and Lin, D. 2018. RIPA: Real-time image privacy alert system. In *CIC 2018*.

Cruz, R. M.; Sabourin, R.; Cavalcanti, G. D.; and Ren, T. I. 2015. META-DES: A dynamic ensemble selection framework using meta-learning. *Pattern recognition* .

De Choudhury, M.; Sundaram, H.; Lin, Y.-R.; John, A.; and Seligmann, D. D. 2009. Connecting content to community in social media via image content, user tags and user communication. In *ICME*. IEEE.

Desai, S.; Caragea, C.; and Li, J. J. 2020. Detecting Perceived Emotions in Hurricane Disasters. In *ACL*. Online.

Kurtan, A. C.; and Yolum, P. 2018. PELTE: Privacy estimation of images from tags. In *AAMAS*.

Li, Y.; Troutman, W.; Knijnenburg, B. P.; and Caine, K. 2018. Human perceptions of sensitive content in photos. In *CVPR Workshops*.

Li, Y.; Vishwamitra, N.; Hu, H.; and Caine, K. 2020. Towards a taxonomy of content sensitivity and sharing preferences for photos. In *CHI*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*. Springer.

Lipford, H. R.; Besmer, A.; and Watson, J. 2008. Understanding Privacy Settings in Facebook with an Audience View. *UPSEC* .

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*.

Orekondy, T.; Schiele, B.; and Fritz, M. 2017. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *ICCV*.

Pleiss, G.; Zhang, T.; Elenberg, E.; and Weinberger, K. Q. 2020. Identifying Mislabeled Data using the Area Under the Margin Ranking. In *NeurIPS*.

Song, X.; Wang, X.; Nie, L.; He, X.; Chen, Z.; and Liu, W. 2018. A personal privacy preserving framework: I let you know who can see what. In *SIGIR*.

Spyromitros-Xioufis, E.; Papadopoulos, S.; Popescu, A.; and Kompatsiaris, Y. 2016. Personalized privacy-aware image classification. In *ICMR*.

Squicciarini, A.; Caragea, C.; and Balakavi, R. 2017. Toward automated online photo privacy. *ACM Transactions on the Web (TWEB)* 11(1): 1–29.

Sundaram, H.; Xie, L.; De Choudhury, M.; Lin, Y.-R.; and Natsev, A. 2012. Multimedia semantics: Interactions between content and community. *Proceedings of the IEEE* 100(9): 2737–2758.

Tonge, A.; and Caragea, C. 2019. Dynamic deep multimodal fusion for image privacy prediction. In *WWW*.

Tonge, A.; and Caragea, C. 2020. Image privacy prediction using deep neural networks. *TWEB* .

Tonge, A.; Caragea, C.; and Squicciarini, A. 2018. Uncovering scene context for predicting privacy of online shared images. In *AAAI*.

Tran, L.; Kong, D.; Jin, H.; and Liu, J. 2016. Privacy-cnh: A framework to detect photo privacy with convolutional neural network using hierarchical features. In *AAAI*.

Wu, Z.; Wang, Z.; Wang, Z.; and Jin, H. 2018. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *ECCV*.

Yang, G.; Cao, J.; Chen, Z.; Guo, J.; and Li, J. 2020. Graph-Based Neural Networks for Explainable Image Privacy Inference. *Pattern Recognition* .

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *ACL* .

Yu, J.; Zhang, B.; Kuang, Z.; Lin, D.; and Fan, J. 2016. iPrivacy: image privacy protection by identifying sensitive objects via deep multi-task learning. *IEEE Transactions on Information Forensics and Security* .

Zerr, S.; Siersdorfer, S.; and Hare, J. 2012. PicAlert! a system for privacy-aware image classification and retrieval. In *CIKM*.

Zerr, S.; Siersdorfer, S.; Hare, J.; and Demidova, E. 2012. Privacy-aware image classification and search. In *SIGIR*.

Zhao, C.; and Caragea, C. 2021. Knowledge Distillation with BERT for Image Tag-Based Privacy Prediction. In *RANLP*.

Zhong, H.; Squicciarini, A. C.; Miller, D. J.; and Caragea, C. 2017. A Group-Based Personalized Model for Image Privacy Classification and Labeling. In *IJCAI*.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .