Towards Automated Content-based Photo Privacy Control in User-Centered Social Networks

Nishant Vishwamitra
University at Buffalo
Buffalo, New York, USA
nvishwam@buffalo.edu

Yifang Li
Clemson University
Clemson, South Carolina, USA
yifang2@g.clemson.edu

Hongxin Hu
University at Buffalo
Buffalo, New York, USA
nvishwam@buffalo.edu

Kelly Caine
Clemson University
Clemson, South Carolina, USA
caine@g.clemson.edu

Long Cheng
Clemson University
Clemson, South Carolina, USA
lcheng2@clemson.edu

Ziming Zhao
University at Buffalo
Buffalo, New York, USA
zimingzh@buffalo.edu

Gail-Joon Ahn
Arizona State University
Tempe, Arizona, USA
gahn@asu.edu

ABSTRACT

A large number of photos shared online often contain private user information, which can cause serious privacy breaches when viewed by unauthorized users. Thus, there is a need for more efficient privacy control that requires automatic detection of users' private photos. However, the automatic detection of users' private photos is a challenging task, since different users may have different privacy concerns and a generalized one-size-fits-all approach for private photo detection would not be suitable for most users. *User-specific* detection of private photos should, therefore, be investigated. Furthermore, for effective privacy control, the exact sensitive regions in private photos need to be pinpointed, so that sensitive content can be protected via different privacy control methods. In this paper, we propose a novel system, `AutoPri`, to enable automatic and user-specific content-based photo privacy control in online social networks. We collect a large dataset of 31, 566 private and public photos from real-world users and present important observations on photo privacy concerns. Our system can automatically detect private photos in a user-specific manner using a detection model based on a multimodal variational autoencoder and pinpoint sensitive regions in private photos with an explainable deep learning-based approach. Our evaluations show that `AutoPri` can effectively determine user-specific private photos with high accuracy (94.32%) and pinpoint exact sensitive regions in them to enable effective privacy control in user-centered online social networks.

CODASPY '22, April 24–27, 2022, Baltimore, MD, USA

ACM ISBN 978-1-4503-9220-4/22/04…$15.00
https://doi.org/10.1145/3508398.3511517

CCS CONCEPTS

• **Security and privacy** → **Privacy protections**; *Usability in security and privacy*.

KEYWORDS

privacy control;social media;deep learning

ACM Reference Format:
Nishant Vishwamitra, Yifang Li, Hongxin Hu, Kelly Caine, Long Cheng, Ziming Zhao, and Gail-Joon Ahn. 2022. Towards Automated Content-based Photo Privacy Control in User-Centered Social Networks. In *Proceedings of the Twelveth ACM Conference on Data and Application Security and Privacy (CODASPY '22), April 24–27, 2022, Baltimore, MD, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3508398.3511517

1 INTRODUCTION

Online Social Networks (OSNs) play an important role in the way Internet users interact with each other to share information. For example, Facebook, the most popular social networking platform, has been used by 79% of Internet users for information sharing [2]. Photo sharing in OSNs comprises of a large portion of the overall content shared in OSNs. An increase in the number of photos shared among users of OSNs has been reported recently [16]. However, a serious implication of this trend is the inadvertent leakage of private information in photos shared by OSN users [9, 27, 40].

Today's OSNs are platforms of massive user interactions, and users by nature are very individualistic. User activities in OSNs are also very individualistic in nature, which makes OSNs inherently *user-centered* systems [29, 36]. Since user behavior cannot be generalized, what constitutes a private photo also cannot be generalized. For example, one user *Alice* would like to publicly share photos depicting herself drinking wine, but another user *Bob* would want to keep such kinds of photos private. Another example of user-specific privacy is the 'bathroom selfie' phenomenon [3]. Although many users might consider the bathroom as a very private space, this recent phenomenon could imply that some users would like to share bathroom photos publicly. Even though some recent studies

attempt to address the photo privacy detection problem, they approach it in a generalized manner. For example, the classifier in [45] learns the privacy level of object classes for all users in OSN in a generalized manner (for example, a cell phone is determined as a private object class for *all* users, just because users may generally consider it private). However, a straightforward photo privacy detection system that simply predicts such a ubiquitous object like a cell phone as a private item for all users would be undesirable to many OSN users. Therefore, a systematic effort is needed in formulating privacy systems that can determine private photos in a *user-specific* manner.

In the online photo sharing domain, information in a photo is a fundamental element at play in privacy [12] and can be interpreted as photo content or elements. Therefore, the content in photos could play a vital role in influencing users' sharing decisions. For example, consider a photo of a user *Alice* depicting her in a bar is uploaded to Facebook by one of her friends. *Alice* is concerned about sharing her identity (such as face and body) in the photo and would therefore like to keep the photo private. Also, *Alice* would not like to share publicly the photo of herself along with alcohol (indicated by a glass of wine). Various information content, such as background, user activities, and objects, can influence users' overall sharing decisions [24]. Some recent studies [27, 32, 35] have discussed the need for content-based privacy in photo sharing. However, a major limitation of these studies is that they do not provide any strategies to *determine* users' sensitive content items. Therefore, there is a need for a system that can infer users' sensitive content items in their photos.

OSN users would generally be aware of their own privacy concerns, but may not be aware of other users' privacy concerns in photo sharing. For example, let's say a user publicly uploads a photo containing her/his friends as well as sensitive content items of her/his friends, due to which her/his friends' privacy is compromised. A reason for this could be that the user is unaware of her/his friends' sensitive content items. An effective strategy for protecting the privacy of users is by controlling the visibility of such sensitive content items in photos. For example, Ilia et al. [27] present a strategy that gives users control over the visibility of their *faces* only in OSNs. However, such strategies do not enable automatic privacy control of all sensitive content items. Thus, *automatic* techniques for determining and controlling sensitive content items for users must be further explored.

In this paper, we propose a novel system, `AutoPri`, which can automatically detect private photos in a user-specific manner, and enable content-based photo privacy control in online photo sharing. `AutoPri` uses a multimodal detection model to detect private photos in a user specific manner and enables content-based photo privacy control leveraging an explainable machine learning technique. Our detection model learns the joint representation of users' photos and their privacy labels (i.e., private or public) using multimodal learning with variational autoencoders. Then, our system integrates an explainable learning-based approach to pinpoint the exact sensitive regions of a private photo to enable effective privacy control of these sensitive content items. To the best of our knowledge, `AutoPri` is the *first* system that can automatically detect users' private photos in a *user-specific manner* and enable content-based photo privacy control.

Our main contributions are summarized as follows:

- **Data Collection and Content Privacy Analysis**. We collect 31,566 photos directly from 303 OSN users with their own privacy concerns. We analyze the content items in the collected photos with respect to users' categorization of the photo as private vs. public and outline insights into sharing patterns of contributors to our dataset. Our analysis of user-specific content privacy lays down important groundwork for building content-based photo privacy control systems.
- **Automatic, User-specific Detection and Content-based Privacy Control**. Our work uses multimodal variational autoencoders (MVAE) [30] to automatically detect users' private photos in a user-specific manner. We further use our explainable learning-based model to pinpoint exact sensitive regions of private photo to enable effective privacy control of these sensitive content items.
- **System Evaluation**. We evaluate the effectiveness of our system based on the dataset that we collected from real-world OSN users. Our evaluation results show that our system is able to accurately determine private photos of users in a user-specific manner with a high average accuracy of 94.32% (along with a precision and a recall of 94.23% and 94.73%, respectively). Our evaluation results based on fidelity tests also demonstrate our explainable model can pinpoint exact sensitive regions of private photos.

The rest of the paper is organized as follows. In Section 2, we introduce the threat model and scope of our work. In Section 3, we discuss our dataset collection strategy. In Section 4, we investigate the importance of content-based, user-specific privacy in photo sharing in OSNs. We articulate our proposed `AutoPri` system, which consists of multimodal learning-based private photo detection model, privacy control of sensitive content using our explainable learning-based model, and the end-to-end flow of our system in Section 5. The details about our system implementation and experimental results are described in Section 6. We discuss limitations and future enhancements of our system in Section 7. We overview the related work in Section 8 and Section 9 concludes our work.

2 THREAT MODEL AND SCOPE

Users. In this work, we consider two types of OSN users: 1) a *photo uploader* is a user who uploads a photo to the OSN; and 2) a *content owner* is a user who is identified in a photo and owns content items in the photo (e.g., their face or possession appears in the photo).

Sensitive Content. In photo sharing, various information content, such as background (e.g., bar, bedroom), user activities (e.g., drinking alcohol) and objects (e.g., container of alcohol, laptop screen) maybe considered sensitive [24] in a private photo. In this work, such information is referred to as *content items* throughout the paper. We assume that a photo is shared privately because it is sensitive to users and do not consider other reasons for sharing privately such as poor quality of the photo or the photo not expected to be of interest to its potential audience.

Threat Model. In this work, we consider the scenario where a photo uploader uploads a photo depicting content owners and their sensitive content, which leads to inadvertent exposure of the sensitive content items to unauthorized viewers. The affected users are

the content owners identified in the photo. We consider any party with access to the original photo as a potential invader of photo privacy. For example, a photo of Jack at a private event depicting sensitive content items, such as a glass of wine, is uploaded online by one of Jack's friends. In this scenario, the affected user is Jack and the sensitive content items are the glass of wine and Jack's identity (such as face and body). All the photo uploader's friends are invaders of Jack's photo privacy. Such inadvertent privacy disclosures can have serious consequences, such as Jack losing his job [1].

Problem Scope. In this work, our goal is to predict the private photos of users to a high degree of accuracy in a user-specific manner, and pinpoint the sensitive regions of a private photo, in order to enable their fine-grained privacy control. Our system is only applicable to users who are a part of the OSN and users who are not a part of the OSN are considered out of scope. In our work, we do not study which content privacy control techniques (such as obfuscation techniques [21, 33] and encryption techniques [22, 42]) for protecting sensitive content items in visual media are best suited, but rather focus on identifying exact sensitive content items of private photos. However, we note that there are emerging studies in the field of human computer interaction that provide excellent guidance about this topic [20, 33]. We do not study the usability of photos after privacy control of sensitive content items in our system. However, there is a large body of existing research that suggests privacy control through obfuscation of sensitive content items in photos is acceptable (further discussed in Section 7). In addition, we assume all OSN providers are trustworthy. Thus, inference attacks by an insider who has access to users' sensitive content items are considered out of scope.

3 DATASET COLLECTION

To train our machine learning model, we need a large dataset that realistically reflects content-based privacy concerns from OSN users. To support automatic content-based photo privacy control, we need a set of photos that contain both photos that users consider private and those that users consider public. But the major challenge in compiling such a dataset is that it is hard to collect users' *private* photos from OSN platforms. Although existing OSNs like Facebook and Instagram provide developer APIs to third parties, these APIs can only be used to collect users' public photos. Some previous work, such as PicAlert [46], attempted to mitigate this problem by re-labeling publicly available photos as private and public using human raters who categorized photos. A major problem with this approach is that it does not capture the original photo owners' own privacy concerns regarding their photos. Furthermore the PicAlert dataset contains only photos that were originally publicly shared (via Flickr). Therefore, the original photo owners did not consider them private. Having external raters re-label them as *private* and *public* may not accurately reflect people's conceptions of private vs. public. For example, the main distinction between the private vs. public photos in PicAlert is whether they were taken indoors vs. outdoors.

In our data collection method, we *directly* ask OSN users for their private and public photos. In this way, we can precisely capture users' own privacy concerns about their photos. Another benefit of this method is that we can collect photos from users who use various OSN platforms (e.g., unlike PicAlert whose users were all from Flickr, our participants used OSNs such as Twitter, Facebook and Instagram; see Section 3, Demographics). This enables us to reinforce platform-agnostic nature of our design. The steps below elaborate the method that we developed for our data collection.

1) Methodology. Our data collection task was approved by our institution's IRB prior to commencement. To preserve participant anonymity, we encoded all participant identifiers in our dataset. In addition, all data is securely stored and can only be accessed by authorized researchers from our institution.

Different OSNs provide different privacy settings for photo sharing. For example, Facebook allows photo sharing with the public (anyone on or off Facebook), friends only, specific friends, or only "yourself". Flickr allows photo sharing with the public, only friends, only family, friends and family, or only "yourself". Similarly other OSNs, such as Instagram, have different privacy settings for photo sharing. Since there are no standard photo sharing privacy settings across OSNs, we decided to categorize the photos in our data collection into two categories, *public* and *private*. In our data collection, the photos shared with the public in current OSNs are mapped as *public* and the photos shared with friends, family, and yourself are mapped as *private*. Note that our system can be generalized to use more granular privacy categories. More details regarding privacy categories are discussed in Section 7.

Participants who volunteered to contribute data to our dataset were provided access to a web application hosted on our server. We first collected demographic information from participants. Next, we provided two boxes titled *public* and *private*, where participants could either drag and drop their photos or upload photos individually from their machines or cell phones. We asked each participant to upload at least fifty of their own photos in each of the private and public categories. We gave participants one guideline to complete this task: *"Private photos are those that you would not share with anyone except your private circles or yourself in online social networks. The photos that you would share with the public in online social networks are considered as public photos."*. On task completion participants were given a unique alpha-numeric code to prove task completion and receive their reward.

We recruited participants using Amazon Mturk. We placed a restriction that only participants with an approval rating of 90% or higher could participate in our data collection task. We offered a $3 reward for completing the task. The average task completion time was around 25 minutes. At the conclusion of the task, we asked participants to re-identify [1] six randomly selected private and public photos that they uploaded. Overall, 361 participants participated in our data collection task. We excluded all data from any participants who had uploaded the same photo in both the private and public categories and excluded participants who uploaded the same photo more than once in the same category. We did this to avoid overfitting in our dataset. In addition, we excluded those participants who failed to re-identify their private and public photos. Using these exclusion criteria, we excluded 58 participants. Our final dataset

[1] We asked participants to identify again some randomly selected private and public photos.

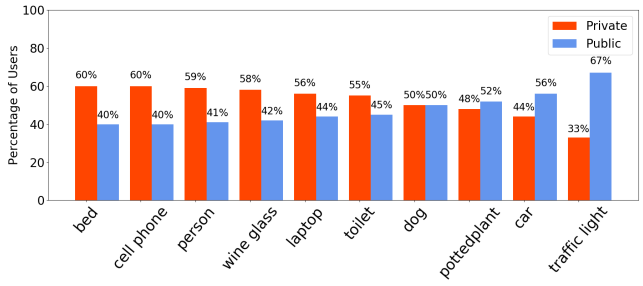


Figure 1: Top 10 objects appearing in the photos with different user privacy concerns in our dataset.

consisted of **31,566** photos from **303** participants, in which 16,058 photos were *private* and 15,508 photos were *public*.

2) Demographics. The demographic information of participants who contributed photos to this dataset is as follows: 53.0% of participants were female, 45.1% were male, and 1.9% chose not to disclose their gender. 18.0% were in the 18-24 age range, 49.8% were in the 25-34 age range, 21.8% were in the 35-44 age range, 6.6% were in the 45-54 age range, and 3.8% were 55 and above. Most participants (90%) reported using Facebook, followed by Instagram (61%), Twitter (50%), Pinterest (40%), and other OSNs (10%).

4 CONTENT PRIVACY CONCERNS ANALYSIS

We present an experiment to understand why we need a system to detect private photos in a user-specific manner. In this experiment, we use all the private and public photos from our dataset to analyze whether users share their photos in a general manner, or in a user-specific manner. For this analysis, we decide to study the objects appearing in the photos, as the objects could give us some indication of why a photo is perceived as private or public by a user. For example, a photo taken in a sensitive situation such as in a bedroom may be private, whereas a photo taken outdoors may be public. We use an existing object detection system, YOLO [7], to detect the objects in a photo. YOLO detects objects from the object categories in the MS COCO dataset [34] that consists of 80 categories of everyday objects [2]. Next, we represent each photo by the object categories that are detected in the photo. Using this setup, we represent the photos as Bags of object categories.

To analyze the different user privacy concerns, we first study the objects that appear in both private and public photos of users. We consider the objects appearing in a private photo as *private* to a user who owns the photo and those appearing in her/his public photos as *public* to the user. We then plot the top 10 objects (the x-axis in Figure 1) appearing in the photos with different user privacy concerns ("Percentage of Users" in Figure 1) in our dataset. From Figure 1, we observe that the users show a very varied perception of privacy towards these objects. For example, the *person* category, which is a very frequently appearing object category (by absolute count), is considered more private (depicted by red bar), but also shared publicly (depicted by blue bar) by a significant portion of users. Photos with objects, such as *bed* and *wine glass*, which we may normally perceive to be more privately-associated, are also found to be perceived publicly by more than 40% users. Objects, such as *toilet*, appearing in very privately perceived situations, are also interestingly, perceived by some users as public (e.g., we find bathroom selfie photos and photos of bathroom decor in public photos), although very low counts of such photos may indicate that users may not like to share such photos. On the other hand, objects that are expected to appear in public setting (such as *car* in outdoor setting) are found to be perceived in a varied manner by users. Thus, any generalized photo privacy control strategy (for example, photos with *toilet* are private) would not be suitable for user-specific privacy control.

[2]For this current study only, we consider the objects belonging to the set of 80 object categories, although our system uses explainable learning-based models to pinpoint such content items automatically, not limited to only object categories in such a dataset.

Table 1: Private and public objects of five randomly-selected users in our dataset.

Identifier	Private Objects	Public Objects
Alice	dining table, sofa, tvmonitor, laptop, wine glass, person	dog, cat, person, pottedplant, car
Bob	person, bottle, backpack	person, car
Carol	cat, person, handbag, pottedplant, diningtable,	bottle, laptop
Dave	bed, person, bottle, pottedplant, tvmonitor	cat, wine glass
Eric	toilet, sofa, car, laptop, person, bottle, pottedplant, diningtable	tvmonitor, cell phone

We next study content privacy concerns considering different users in our dataset and observe that users perceive similar types of photos in very different ways. To illustrate this finding, we depict the most privately and publicly perceived objects of five randomly selected contributors in our dataset, as shown in Table 1. For the sake of anonymity, we refer to those users as *Alice*, *Bob*, *Carol*, *Dave*, and *Eric*, respectively. From Table 1, we can observe that different users have different perceptions of private vs. public regarding the same object. For example, *Eric* perceives 'car' as *private*, whereas 'car' is public for *Alice*. A similar observation can be made regarding 'cat', for *Carol* and *Alice*.

In summary, we find that different users perceive the sensitive content of their photos in very different manners. A generic photo privacy detection approach [45] cannot handle such varied user perceptions towards shared photos. Therefore, we need a system that can automatically determine the sensitive content of shared photos in a *user-specific* manner.

5 SYSTEM DESIGN

In this section, we describe our system, `AutoPri`, for the detection of users' private photos in a user-specific manner and the support of content-based privacy control of the photos. First, we discuss the intuitions behind our system design. We describe how `AutoPri` detects the private or public photos of users in a user-specific manner based on the multimodal learning of users' photos and their privacy labels. Then, we discuss the method of pinpointing the exact sensitive regions of a private photo using our explainable learning-based model. Lastly, we present the end-to-end flow of the `AutoPri` system, in which we elaborate how it would be operational in a potential OSN environment.

5.1 Design Intuitions

5.1.1 Multimodal Learning for User-specific Privacy Detection. The first objective of our system is the automatic detection of private photos of a user in a user-specific manner. However, the detection of private photos in such a user-specific manner is a complex problem, because each user has different privacy requirements that are distinct from other users' privacy requirements. The user-specific nature of this problem requires a model to learn a joint representation of two modalities - user information and photo information. Thus, this problem cannot be solved by straightforward supervised learning techniques [14] as they are not capable of learning joint representations of data with multiple modalities.

Thus, to predict privacy labels (such as "private" or "public") for a user's photo in a user specific manner, we need a model that can learn the joint representation of a user and their photos, which are actually two different modalities of information. For example, consider a user *Alice* who shares photos of herself in a bar with the "private" privacy label. However, another user *John* likes to share his bar photos with the "public" privacy label. Thus, these two users have completely different privacy concerns for the same kind of photo. This is a multimodal problem because there are two modes of information, user information (i.e., *Alice* or *John*) and photo information (i.e., bar photo), which should be used by a detection model in order to make a user-specific prediction.

A major challenge in multimodal learning is how to learn such joint representations. To address the above challenge, we use a multimodal variational autoencoder (MVAE) [30] based model (called "`AutoPri` detection model" in this paper) in our system that can learn the privacy label of a photo in a user specific manner by learning the joint representation of users and their photos. The `AutoPri` detection model learns to combine the two modalities of users and photos by learning their joint representation using a product-of-experts (POE) inference network [23, 44]. Given an inference network for each modality, the POE network learns the joint representation of each of these modalities. Thus, the problem of user and photo modalities in user-specific privacy detection can be addressed by learning their joint representation with MVAEs and the POE inference network.

5.1.2 Explainable Learning for Photo Privacy Control. The second objective of `AutoPri` system is to enable privacy control in photo sharing. To enable privacy control of sensitive content, several effective solutions have been formulated [20, 33]. However, the challenge here is how to know which photo regions are sensitive so that they can be controlled. To control the sensitive regions, we need to know the exact pixels in the photo to apply a suitable control technique. Ideally, we would want the `AutoPri` detection model to tell the photo regions that are responsible for the "private" privacy label for a user's photo. However, this information is not available from a classification model because deep neural networks are normally black-box techniques and it is not possible to know which regions the model focuses on to classify a user's photo as "private". Although we have used the YOLO object detection system for content-based analysis in Section 4, which can produce bounding boxes around detected objects, such state-of-the-art object detection systems have several crucial limitations, due to which they cannot be used in privacy control systems. Firstly, YOLO (and other such object detection systems) cannot distinguish between objects in the same category. For example, a national flag and the confederate flag are detected as the same object category - "flag". Secondly, YOLO trained on MS COCO dataset is limited to only 80 object categories. We found that these categories are *not sufficient* to include a large amount of private content. Thus, we need a method that can pinpoint sensitive regions in a more fine-grained manner and not limited to only categories available in MS COCO or other datasets with a small number of object categories.

Explainable deep learning [39, 47] is a recent branch of deep learning techniques that focuses on generating *explanations* for a model prediction. In computer vision applications, such explanations can pinpoint the regions of photos that are responsible for a prediction. Applying explainable deep learning for private photo detection, explainable models can pinpoint regions of a photo that could be responsible for the "private" label of the photo. For example, Figure 2 (a) depicts a scene with a beer bottle and beer mug, which could be responsible for the "private" label of this image. Figure 2 (b) depicts two pet animals that could be of concern to the photo owner and thus responsible for the "private" label for this photo. Figure 2 (c) depicts a person with his car.

Figure 2: Examples of sensitive items that explainable models can pinpoint in private photos: (a) beer bottle and beer mug; (b) two pet animals; and (c) person with car. (Note that all photos used in this paper are stock photos from Google images, labeled for noncommercial reuse with modification.)

In the `AutoPri` system, the detection model uses a multimodal variational autoencoder to learn the join representation of users and photos as discussed in Section 5.1.1. We use the existing convolutional neural network in the photo encoder of our detection model to generate explanations of privacy labels. Once we generate the explanations, we can apply suitable privacy control techniques to the pinpointed regions to enable fine-grained photo privacy control.

5.1.3 Content-level Photo Privacy Control. There are currently several content-level privacy control techniques including obfuscation [27, 33] and encryption [22, 42] proposed for photo privacy protection. In particular, obfuscation has been demonstrated to be suitable for content-based photo privacy control. Furthermore, researchers are currently focusing on which obfuscation techniques are best suited for privacy protection tasks [21, 33]. For example, several other obfuscation techniques such as avatar and inpainting [33] have been shown to provide a better sense of human contact and offer better adoption willingness than conventional obfuscation techniques such as blurring. Another technique called cartooning [21] has recently emerged as an unobtrusive form of obfuscation, preferred by OSN users. In our work, we focus on the

detection of sensitive content in photos in a user-specific manner, and provide the control of such sensitive content items through the privacy control techniques discussed here, according to the preferences of the OSN users. Choosing the most effective or suitable privacy control techniques is out of our scope. Since there is a rich body of work that discuss in-depth user studies about the usability and willingness of users to adopt these privacy control techniques, we do not conduct such studies about content-level privacy control in our work.

5.2 AutoPri System Design

We first give an overview of AutoPri system that includes a training phase (Figure 3, "Offline Training") to train the AutoPri detection model and an online evaluation phase (Figure 3, "Online Evaluation") to enable privacy control in a user's photo detected as private. Our dataset, which consists of users' photos and their labels, is first fed into the AutoPri model (Figure 3, Step (i)). We represent users and photos as a joint representation and input them into the User Photo Encoder (Figure 3, Step (ii)). The photo labels are fed into the Label Encoder (Figure 3, Step (ii)). In our models, we use deep neural networks for encoding the user photos and the privacy labels. Next, we learn the joint representation of user photos and their privacy labels. We use the product-of-experts (POE) technique to learn the joint distribution of the user photos and their privacy labels (Figure 3, Steps (iii) and (iv)). The POE generates a Latent Representation that represents this joint distribution of user photos and their labels (Figure 3, Step (v)). The Latent Representation is then fed into the User Photo Decoder to reconstruct back the user photo and the Label Decoder to reconstruct the photo label, respectively (Figure 3, Step (vi)). After the training process, the Label Decoder is used to output user-specific predictions (Figure 3, Step (vii)).

The online evaluation phase in Figure 3 represents the processes involved in the evaluation of a new photo when it is uploaded and enabling privacy control if detected as private. When a user's photo is uploaded, the AutoPri system first detects its user-specific privacy label using its detection model (Figure 3, "Private Photo Detection"). If the photo is detected as private, the AutoPri explainable model pinpoints the exact regions that are responsible for the private label (Figure 3, "Pinpoint Sensitive Regions"). Then, based on the explained regions, bounding boxes are generated surrounding those regions. Finally, we apply suitable privacy control techniques to the sensitive regions enclosed by bounding boxes to hide sensitive content in the shared photo (Figure 3, "Privacy Control").

5.2.1 User-specific Photo Privacy Detection. We first encode the user information together with the photo information. We express each user as a one-hot encoding of the user that is encoded into the user's photo as an additional channel, so that the user photo encoder can see both these information in a joint fashion. In the AutoPri detection model, we use a deep CNN to encode this joint information. We express the privacy labels as one-hot encodings and we use a multi-layer perception (MLP) network to encode these labels. Our objective is to learn the joint representation $p_\theta(I, L, z)$ of the two conditionally independent modalities, including user image I and privacy label L, given a common latent variable z (Figure 3, "Latent Representation"), which is factorized as follows.

$$p_\theta(I, L, z) = p(z)p_\theta(I|z)p_\theta(L|z) \tag{1}$$

From Equation 1, we can ignore the missing label (that we want to predict) during the prediction time of the user's photo, thus generating the label from the joint distribution. Next, we approximate the joint posterior $q(z|I, L)$ as a product-of-experts [23]. From more recent work [44], the joint posterior can be approximated as the product of individual posteriors, as depicted in Equation 2.

$$p(z|I, L) = p(z)\tilde{q}(z|I)\tilde{q}(z|L) \tag{2}$$

Thus, we can use a product-of-experts, including a prior expert ($p(z)$) as the approximate distribution of the joint posterior. In Equation 2, $\tilde{q}(z|I)$ and $\tilde{q}(z|L)$ are approximated with neural networks based encoders. Next, to compute the POE, we consider the $p(z)$, $\tilde{q}(z|I)$ and $\tilde{q}(z|L)$ as Gaussian distributions, so that the product of experts is also a Gaussian with mean (μ) and variance (V) computed as follows.

$$\mu = \frac{\mu_I T_I + \mu_L T_L}{T_I + T_L} \tag{3}$$

$$V = \frac{1}{T_I + T_L} \tag{4}$$

Where T_I and T_L are V_I^{-1} and V_L^{-1}, respectively, in Equations 3 and 4. In Figure 3, the user photo encoder and the label encoder produce the mean and variance (Figure 3, Step (iii)), which are then combined to produce the latent representation z (Figure 3, Step (v)), using the POE technique.

The generation of latent representation z is followed by the process of reconstruction with the user photo decoder for reconstructing photos and users, and the label decoder for reconstructing the privacy label (Figure 3, Step (vi)). To train the model, we use the ELBO loss [30], defined for joint representation of user photos and privacy labels as described below.

$$ELBO(X) = \mathbb{E}_{q_\phi(z|X)}[\sum_{x_i \in X} logp_\theta(x_i|z)] - \beta KL[q_\phi(z|X), p(z)] \tag{5}$$

In Equation 5, $X \in \{I, L\}$. The first term in Equation 5 signifies the reconstruction loss of the user photo and the label, and the second term signifies the Kullback-Leibler divergence [30] between the distributions p and q.

5.2.2 Explaining Regions of Privacy in Photos. If the AutoPri detection model detects a user's photo as private, our objective is to enable the privacy control of the sensitive regions of the photo that are responsible for the privacy concern. In our system, we use explainable machine learning techniques [39, 47] to pinpoint these sensitive regions. We use gradient-based explainable technique for CNNs to first generate an "explanation" for the privacy label, based on the feature maps computed in the convolutional layers of the user-image encoder as illustrated in Figure 3 (ii) ("Feature Maps"). This explanation is of the form of an activation map that is activated for the regions of the photo that are most responsible for causing the privacy label (Figure 3, "Explanation Results"). Next, we use activated regions of the activation map to generate bounding boxes that enclose the private regions (Figure 3, "Bounding Boxes"). Photo privacy control techniques that are preferred by a user are then

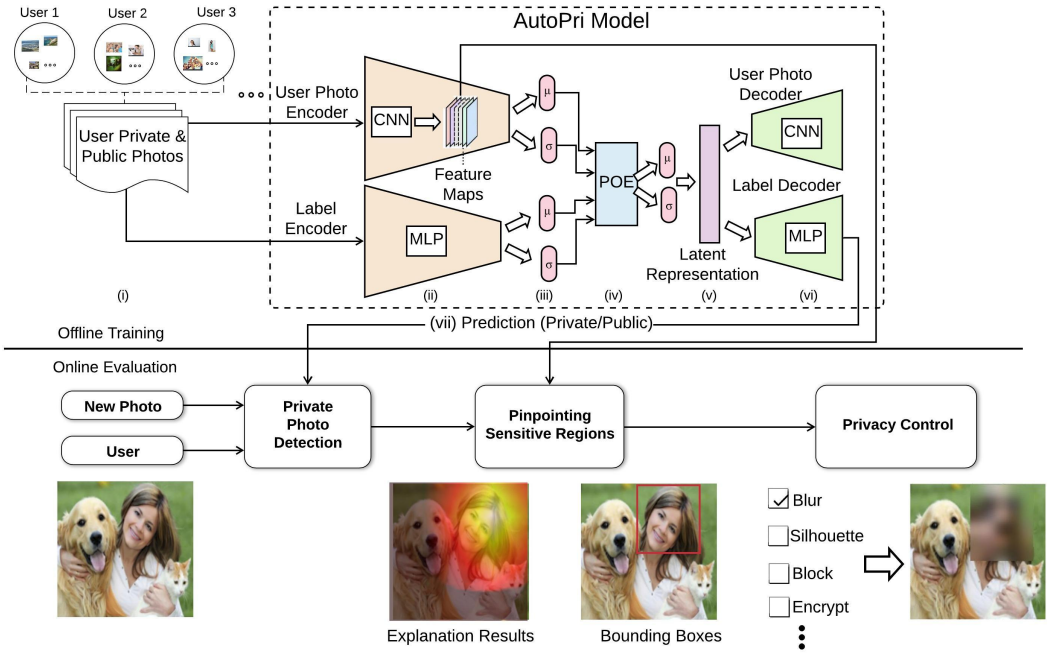


Figure 3: Overview of AutoPri system.

applied to the photo regions enclosed by the bounding boxes so that these photo regions are prevented from being viewed (Figure 3, "Privacy Control").

To compute the activation map, we first compute the gradient of the "private" privacy label ($y^{private}$) with respect to the feature maps A^k of the last convolutional layer of the user photo encoder (Figure 3, "Feature Maps"), i.e., $\frac{\partial y^{private}}{\partial A^k}$.

These gradients are average pooled to obtain importance weights ($a_k^{private}$) as shown in Equation 6.

$$a_k^{private} = \sum_i \sum_j \frac{\partial y^{private}}{\partial A_{ij}^k} \tag{6}$$

To generate the activation map, the importance weights from Equation 6 are combined with the feature maps as shown in Equation 7. We apply the ReLU activation function because we are only interested in the pixels that are most important to the private label. The ReLU [38] operation helps to eliminate the negative values in the activation map.

$$activation\ map = ReLU(\sum_k a_k^{private} A^k) \tag{7}$$

To generate a bounding box, we upsample the activation maps to the photo dimensions and use a segmentation technique. We segment the pixels having a value greater than or equal to 40% of the maximum value in the upsampled activation map. Then, we generate box coordinates that can cover this region. Finally, we use the photo privacy control technique selected by the user to protect this region in the photo.

5.2.3 End-to-end System Flow. We would like to summarize the end-to-end flow of AutoPri as illustrated in Figure 4. The processing of a photo begins when a user uploads it to the OSN (for example, as shown in Figure 4 (i), the user *Alice* uploads a photo containing herself, the user *John* and *John's* dog). This is followed by the identification of all OSN *content owners* in the photo using the face identification technology (as shown in Figure 4 (ii), *Alice* and *John* are detected). Next, AutoPri initiates the following actions according to the *content owners* identified in the photo. First, the AutoPri detection model is used to get the detection score for the new photo with each *content owner* identified in the photo. Second, if the privacy label for the photo is detected as "private" for any of the identified *content owners*, AutoPri's explanation model is used to pinpoint sensitive content items of the photo for these *content owners*. Third, the uploader is asked to control the sensitive regions (as shown in Figure 4 (iii)). Finally, the photo can be shared with the sensitive content regions controlled.

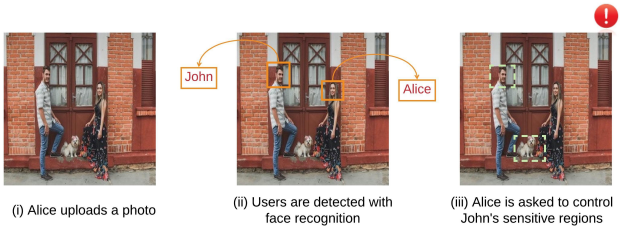


Figure 4: End-to-end flow of AutoPri system.

6 IMPLEMENTATION AND EVALUATION

6.1 System Implementation

The user photo encoder of the AutoPri detection model is a CNN with six convolutional blocks, consisting of convolution, ReLU, and max-pooling layers. The user photo decoder is also a CNN with six convolutional blocks, consisting of transposed convolutions, ReLU, and max-pooling layers. The label encoder and decoder are MLP

networks with fully-connected linear layers and ReLU activations. We use the Binary Cross Entropy [28] loss as the reconstruction loss for both the user photo and the label. During the training time, we let the model train with the training dataset for certain number of iterations, followed by letting the model train on random signals as privacy labels for the next number of iterations to enable multimodal training. Our models have been developed using the PyTorch [37] framework, on computing platforms consisting of NVIDIA V100 GPUs and Intel Xeon multi-core CPUs.

6.2 System Evaluation

In this section, we perform experiments to evaluate our system from several different perspectives. The major goals of our evaluation are summarized as follows.

- Evaluating the effectiveness of `AutoPri` detection model in accurately detecting private and public photos for different users in our dataset (Section 6.2.1).
- Evaluating the effectiveness of `AutoPri` detection model in the *user-specific* detection of users' private photos (Section 6.2.2).
- Evaluating the effectiveness of `AutoPri` explainable model in pinpointing the exact sensitive content items of private photos (Section 6.2.3).
- Evaluating the effectiveness of the system to prevent privacy invasion from potential users' perspective by conducting an experiment with OSN users (Section 6.2.4).

6.2.1 Effectiveness Evaluation of `AutoPri` *Detection Model.* To evaluate the effectiveness of `AutoPri` detection model in the detection of private and public photos of users in a user-specific manner, we randomly select 80 percent of our dataset for training (with 5-fold cross validation) and 20 percent of the dataset for testing, and we run our system on photos in our test dataset. We perform the Receiver Operating Characteristics (ROC) [17] analysis of our model for user-specific detection of private and public photos. The ROC analysis provides a means of reviewing the performance of a model in terms of the trade-off between False Positive Rate (FPR) and True Positive Rate (TPR) in the predictions. The ROC plot of the `AutoPri` detection model is depicted in Figure 5. Overall, our detection model achieves an overall precision and recall of 94.23% and 94.73%, respectively, on the test dataset, with an overall accuracy of 94.32%. From Figure 5, the area under the curve (AUC) for `AutoPri` detection model is 0.99, which indicates a good balance of false positives and false negatives.

Next, we investigate the false positives in this experiment. Among the 220 false positives out of 7401 test photos (2.9%), it was found that these photos were very similar to photos by same user shared as public, due to which the model is not able to distinguish between them. We may note that this is not due to the detection model not being able to distinguish accurately the difference in the privacy concerns regarding the photos, but it could be due to the users' own mis-identification of private and public photos.

6.2.2 Evaluation of User-specific Detection. The `AutoPri` detection model supports the *user-specific* detection of private photos. Our objective in this experiment is to evaluate whether `AutoPri` detection model is indeed generating user-specific detection scores of shared photos.

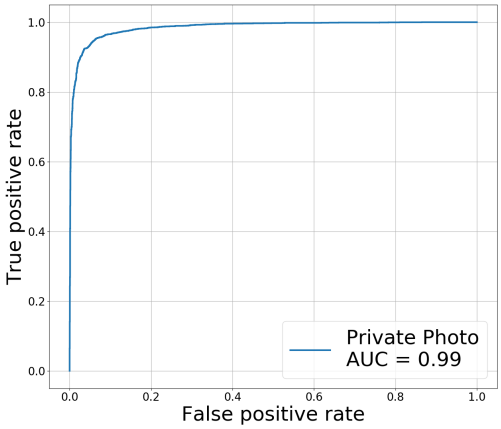


Figure 5: ROC evaluation of `AutoPri` detection model.

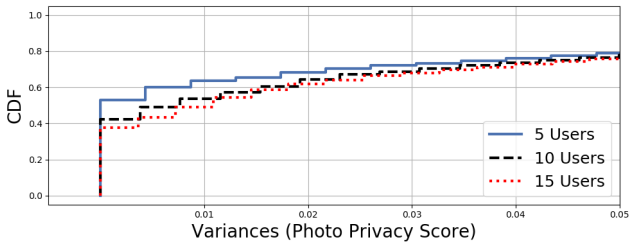


Figure 6: Evaluation of user-specificity of `AutoPri` detection model. Figures depict the variance in detection scores for 5, 10 and 15 users for each photo.

Users' could have different privacy concerns regarding the privacy of photos. Thus, to ensure user-specific detection, our model must generate different detection scores for different users. In this experiment, we study the variance of the detection score for the same photo considering different users in our test dataset. Variance is a measure of the degree of variation in an observation. In the user-specific photo privacy detection, variance would indicate the degree of variation in the detection scores for different users. For example, a photo could be private to some users but could be public for other users. Thus, higher variance indicates a large variation (i.e, low homogeneity) in detection scores and lower variance indicates a low variation (i.e, high homogeneity) in detection scores.

In this experiment, we first randomly choose 1000 photos from our dataset. For each of these photos, we randomly select 5, 10 and 15 users from our data collection (Section 3) and assume they are the owners of the photo. We run the `AutoPri` detection model for each number of users. Next, we compute the variance of the detection scores for each photo for the 5, 10 and 15 randomly selected users. We depict the results of this experiment in Figure 6.

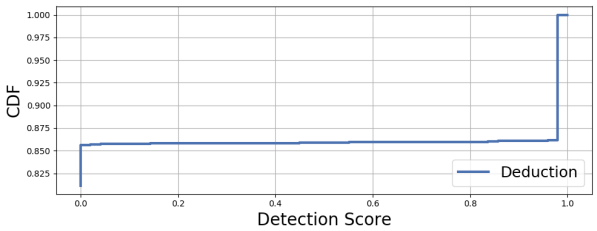
From Figure 6, the first observation is that some photos may have very low variance for many users. For example, from Figure 6, around 56% of photos for 5 users case, around 42% of photos for 10 users case, and around 38% of photos for 15 users case have very

low and nearly 0 variance. This could indicate that for many photos, most users may have the same privacy concern. For example, we observed that photos depicting only outdoor scenes do not have any private content for any user, as a result of which these may have similar privacy concerns.

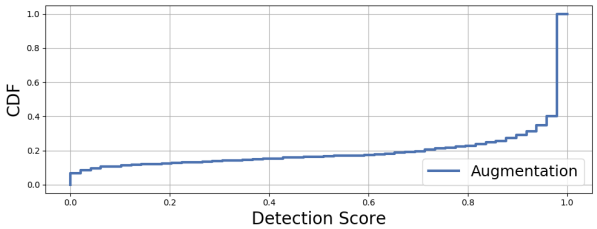
From Figure 6, the second important observation is that many photos may have high variance for many users. For example, from Figure 6 around 44% of photos for 5 users case, around 58% of photos for 10 users case, and around 62% of photos for 15 users case have higher variance for the photos. We observed that these photos depicted very specific scenes, for example, private occasions. These also indicate that many users have user-specific privacy concerns for many of their photos.

Another observation from Figure 6 is that as the number of users for a photo increases, the variance in the detection scores also shows a similar increase. This observation may indicate that as the number of users for a photo increases, the variation in the detection scores also increases due to more variations in the privacy concerns of many users towards the photo.

6.2.3 Explainable Model Evaluation. The `AutoPri` privacy control strategy is based on explainable machine learning to pinpoint the sensitive regions that are responsible for photo privacy. Our objective in this experiment is to evaluate whether `AutoPri` explainable model is indeed pinpointing effectively the sensitive content items in private photos. We evaluate the effectiveness of the `AutoPri` explainable model using fidelity metrics (deduction and augmentation) [19]. The objective of our evaluation is to evaluate the *correctness* of the sensitive content items that are pinpointed by our explainable model, as these content items are responsible for the private label of a photo for a specific user. These content items are also crucial as they are used in the privacy control strategy (Section 5.2.3) in `AutoPri`.



(a) CDF of PCR after deducting sensitive content items from private photos.



(b) CDF of PCR after augmenting random public photos with sensitive content items from private photos.

Figure 7: Evaluation of `AutoPri` explainable model.

We denote an original private photo as x and the sensitive content items pinpointed by `AutoPri` explainable model in x as V_x. We conduct two tests to validate the correctness of sensitive content items, listed below.

- **Deduction Test.** If the sensitive content items V_x pinpointed in a private photo are correct, then removal of V_x from x should lead to the re-classification of x as public photo. In this test, we use blocking to remove the content items of x pinpointed as sensitive by our explainable model and re-classify the photo x using the `AutoPri` detection model.
- **Augmentation Test.** If the sensitive content items V_x pinpointed in a private photo are correct, then adding V_x to a public photo x should lead to the re-classification of x as a private photo. In this test, we randomly select a public photo and add V_x from a randomly selected private photo and re-classify the new photo using the `AutoPri` detection model.

For each of the above two tests, we create two samples of each photo in our test dataset. Then, we input each sample into our detection model and examine the Positive Classification Rate (PCR). In this method, we examine the ratio of photos that are re-classified with their original label. For example, in the deduction test, we would ideally expect a low PCR, since the removal of sensitive content items from a private photo must render the photo as public. Similarly, we would ideally expect a high PCR for the augmentation test.

In the deduction test, from Figure 7a, it can be observed that most of the private photos are reported as public after the deduction of the sensitive content items. After removing the sensitive content items pinpointed by the explanation model from private photos, an overall drop of PCR to 15.87% was observed. Considering that our detection model achieves an accuracy of 94.32%, the drop in PCR to 15.87% may indicate that the sensitive content items pinpointed by our explainable model are the regions in photos responsible for high privacy level of the photos.

In the augmentation test, from Figure 7b, it can be observed that most of the randomly selected public photos are reported as private after augmentation of the sensitive content items in them. After adding only the sensitive content items pinpointed by the explainable model to randomly selected public photos, a PCR of 82.7% was observed. Considering that our detection model achieves an accuracy of 94.32%, the rise in PCR to 82.7% may indicate that even the sensitive content items pinpointed by our explainable model alone are highly responsible for high privacy level of the photos.

Figure 8: Explanations of user-specific detection. Figures (i) and (ii) depict the explanation and the bounding boxes generated for one user. Figures (iii) and (iv) depict the explanation and bounding boxes generated for a different user.

We used our explainable model to visualize some examples of the different content items of a photo, but evaluated for randomly selected and different users from our dataset. Figure 8 shows a photo

that is evaluated for two randomly selected users. From Figures 8 (i) and (iii), it can be observed that the explanation model focuses on different regions to infer the privacy scores for different users. For example, in Figure 8 (i), the focus is on the person (indicated by brighter, yellow color of heatmap), but in Figure 8 (iii), the focus is on the cat. Thus, the sensitive content items pinpointed by our system are different for different users based on their different privacy concerns.

6.2.4 Experiment With Potential Users. To evaluate AutoPri's ability to prevent the invasion of privacy from potential users' perspective, we have conducted a preliminary online experiment with 57 of the same participants who participated in the dataset collection task (Section 3).

Methodology. In this experiment, we recruited 82 of the Amazon Mturk participants who had participated in our data collection task, 57 of whom completed the task. Only the Amazon Mturk participants who participated in our earlier task were considered eligible for this experiment. Each participant received compensation of $1 for their participation (the average task completion time was around 2 minutes 30 seconds). The entire experimental protocol was approved by our institution's IRB.

We studied the ability of our system to prevent the *invasion of privacy* [8] of shared photos. The experimental design is within-subject, meaning all participants received all experimental conditions. We chose blurring as the privacy control technique for this experiment, as it is used in many current applications such as YouTube [10] and Google Street View [18], due to which it could be most familiar to a general user. Note that our objective in this experiment is *not* to study about blurring as a photo privacy control technique, but to study the ability of our system in reducing invasion of privacy in specific users' own photos. There are two independent variables, each with two levels. The independent variables are photo type (private or public) and content item blurring (altered or unaltered). When crossed, this results in four experimental conditions: (1) private photo with blurring; (2) private, unaltered photo; (3) public photo with blurring; and (4) public, unaltered photo (see Figure 9).

Our experiment procedure is as follows. First, we randomly selected one private photo and one public photo from the set of photos each participant uploaded during the data collection task to use as stimuli for the experiment. Each participant was shown four stimuli photos, one for each experimental condition, in *random* order. The four experimental conditions are: (a) participant's private photo with sensitive content items which were pinpointed by AutoPri and blurred out (called *altered private photo*); (b) unaltered version of the participant's private photo (called *unaltered private photo*); (c) participant's public photo with some content items blurred (called *altered public photo*); (d) unaltered version of the participant's public photo (called *unaltered public photo*) (Figure 9). Since, by definition, there should be no sensitive content in photos identified as public, we randomly selected content items to blur. For each of the four photo stimuli we presented to participants, we asked participants to respond to the statement *"I feel my privacy can be compromised because sensitive content can be learned from this photo"* [8], using a five point Likert scale (1-Strongly disagree, 2-Disagree, 3-Neither agree nor disagree, 4-Agree, 5-Strongly agree).

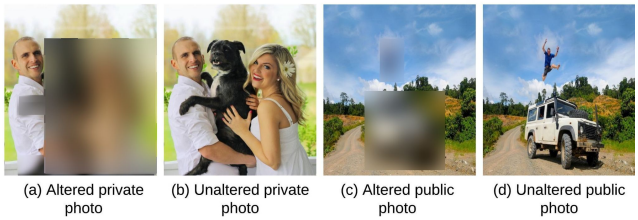


Figure 9: Example photos demonstrating the four experimental conditions: (a) altered private photo, (b) unaltered private photo, (c) altered public photo, and (d) unaltered public photo.

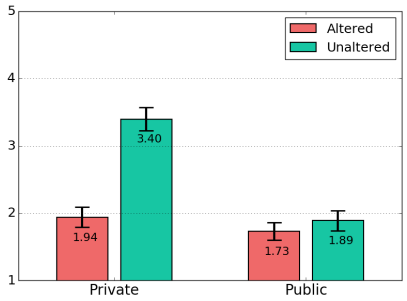


Figure 10: Mean (SE) agreement with the statement "I feel my privacy can be compromised because sensitive content can be learned from this photo".

After participants completed the experiment, we asked participants to view three randomly selected private and three randomly selected public photos from the original set of photos they had uploaded, and asked them to re-identify their private and public photos. We excluded 25 participants who failed to correctly re-identify their private or public photos for a final sample size of 57.

Results. Figure 10 depicts the mean (Standard Error (SE)) of participant responses to the statement, "I feel my privacy can be compromised because sensitive content can be learned from this photo." Overall, we see that, as expected, unaltered photos pose more perceived risk of privacy invasion than altered photos, suggesting participants understood and completed the task correctly. We also see that both unaltered and altered public photos, and altered private photos pose very little perceived risk of privacy invasion (with means between 1 and 2, indicating disagreement to strong disagreement with the statement, "I feel my privacy can be compromised because sensitive content can be learned from this photo."). The only experimental category of photos to result in agreement (M=3.40, corresponding to between "neither agree nor disagree" to "agree") is unaltered private photos.

To examine this finding in detail, we conducted dependent t-tests (paired t-tests) [25]. While the response categories to the statement are ordinal, we treat them as interval for the purposes of analysis [31, 43]. Again, looking at Figure 10, for private photos we see that participants' agreement with the statement "I feel my privacy can be compromised because sensitive content can be learned from this photo" was much lower for the altered version of the private photo (Mean 1.94, SE 0.15) compared to the unaltered private photo (Mean 3.40, SE 0.17) ($p < .001$, $r = .872$) [3]. For public photos, we also see

[3] A small p-value ($p < .05$) indicates strong evidence against null hypothesis. An effect size (r = 0.8) indicates a large effect size. An effect size (r = 0.2) indicates a small effect size.

a difference between participants' agreement with the statement for the altered public photo (Mean 1.73, SE 0.13) and the unaltered public photo (Mean 1.89, SE 0.15) ($p < .01$, $r = .359$), however the effect size was relatively small. The small effect size could reflect participants' perception that public photos may not pose as much of a privacy risk overall. Together, these findings suggests that the system, by identifying and blurring sensitive content items in private photos, reduces participants perception that their privacy can be compromised because of the sensitive information in a photo.

One limitation of our experiment is that we showed the same private and public photos for both the altered and unaltered conditions. However, this limitation is mitigated because we randomized the order in which participants viewed altered vs. unaltered photos (i.e., some saw altered first, others saw unaltered first). The randomized ordering allows us to limit demand effects, as users also see the blurred version of their own public photo in a random order (and not just blurred version of their own private photo). The second limitation could be that we studied the invasion of privacy based on one question. However, this helps us to avoid directly asking about the privacy of a photo, which is an effect that has been shown to result in biased responses [11]. To limit demand effects, we did not elicit participant responses at a content level, although such a study, when designed carefully, could lead to interesting insights about content-level photo privacy.

7 DISCUSSION

In this section, we discuss some limitations and potential enhancements of our work. It should be noted that this work represents the *first* step towards the design of a content-based automatic privacy control system for photo sharing in OSNs in a *user-specific* fashion.

Dataset Limitations. A limitation of our data collection method would be that some very privacy sensitive participants may not have been willing to share photos containing extremely sensitive content. Also, in our data collection method, we used private and public as the privacy settings. As part of our future work, we plan to use more granular categories as the privacy settings in our method, such as friends, close friends, family, colleagues etc, or user-defined categories.

Deployment. We present details about the deployment of `AutoPri` in practical systems. Our system can be integrated in the current OSNs. Users can provide `AutoPri` access to their photos in OSNs, which would be used to train our model. Deep learning-based models have been successfully deployed in various current OSNs, such as Facebook [5] and Instagram [6]. For example, upon uploading a new photo, Facebook suggests tagging users present in the photo. Many current OSNs, such as Facebook, also use face detection technology to improve the social media experience of their users [4]. In a deployment scenario, `AutoPri` can use the existing face detection algorithms available in current OSNs to identify users in photos, and then suggest sharing settings, along with the privacy control of sensitive regions by pinpointing them in a new photo.

There are two interesting scenarios, which also need to be discussed in the adoption of `AutoPri` in an existing OSN. First, the user may be a new OSN user. As a result, the user would not have a significant number of shared photos in the OSN. In `AutoPri`, we propose that we could help new users by initially identifying sensitive content items based on general privacy concerns of other users. Then, subsequent suggestions can be made more specific to the user as more individual sharing data is accumulated. A second scenario involves a user who has an unbalanced number of private and public photos. We propose to augment the dataset of such a user with external photos, which reflects users' preferences. For example, in `AutoPri`, a user's close friends photos could be used to augment the user's dataset and mitigate the unbalanced data problem.

8 RELATED WORK

Most of the existing works that study privacy in online photo sharing only discusses sharing at the resource level (entire photo) [9, 26]. However, sharing at the content-level is an intuitive, human-centered solution for protecting photo privacy [13], because it avoids forcing users to take an all-or-nothing approach of either sharing or withholding entire photos. Although several other works such as [15, 40, 41] also discuss important insights about photo privacy, *content-level* privacy is not discussed in these works. Recently, some emerging studies [27, 32] have attempted to address the content-level photo privacy. For example, the importance of content privacy is discussed in [32]. However, this work does not provide effective solutions to address content-based privacy leakage. The work in Face/off [27] provides an investigation into the sharing behaviors of users in OSNs. The study reveals that most users are not aware of sharing their photos with individuals unknown to them. Several privacy leakage scenarios are discussed to express the importance of privacy in photo sharing over OSNs. However, there is no solution provided towards the protection of sensitive content of a user, since the Face/off model can only handle privacy protection towards an OSN user's face, although *face* is just one of the many content items that can be sensitive to a user [33].

The importance of user-specific determination of sensitive content has been discussed in recent works [41, 45]. However, previous works only address online photo privacy protection problem in a generalized manner. The iPrivacy system [45] presents a method to address automatic sensitive object class detection by training a tree classifier to predict the sensitive object classes from a large number of photos. Although this work discusses protection of sensitive object classes, a crucial drawback that impacts its practicality is that the sensitive object classes are determined in general for all users in OSNs. Other studies such as [41] are also limited because they do not address user-specific privacy concerns. Therefore, the general models of photo privacy protection are impractical for handling *user-specific* privacy concerns in OSNs.

9 CONCLUSION

In this paper, we have proposed `AutoPri`, a system for the automatic detection of users' private photos in a user-specific manner and the effective privacy control of sensitive content items. We have collected a large and realistic dataset of 31,566 photos from 303 OSN users with their own privacy concerns and discuss how users may have specific privacy concerns regarding their sensitive content items. We have further discussed the need for automatic and user-specific content-based detection for users sharing photos in OSNs. We have presented our system `AutoPri`, that consists of a detection model with a multimodal variational autoencoder and

an explainable deep learning-based model for automatic and user-specific content-based photo privacy control. The evaluation of our system demonstrates the effectiveness of AutoPri in detecting user-specific private photos with a high accuracy and with low performance overhead. Our experimental results also demonstrate our explainable model can accurately pinpoint the sensitive regions in private photos to enable effective privacy control.

ACKNOWLEDGMENT

This work is supported in part by the National Science Foundation (NSF) under the Grant No. 2129164, 2114982, 2031002 and 2120369.

REFERENCES

[1] 2011. Teacher sacked for posting picture of herself holding glass of wine and mug of beer on Facebook. https://www.dailymail.co.uk/news/article-1354515/Teacher-sacked-posting-picture-holding-glass-wine-mug-beer-Facebook.html.
[2] 2016. Social Media Update 2016 - Pew Research Center. https://www.pewresearch.org/internet/2016/11/11/social-media-update-2016/.
[3] 2017. Bathroom Selfie Phenomenon. https://www.theguardian.com/media/shortcuts/2017/jan/24/finished-in-there-yet-how-bathroom-selfie-became-huge.
[4] 2020. Face Plus Plus. http://www.faceplusplus.com.
[5] 2020. Facebook. https://www.facebook.com.
[6] 2020. Instagram. https://www.instagram.com/.
[7] 2020. YOLO. https://pjreddie.com/darknet/yolo/.
[8] Ramakrishna Ayyagari, Varun Grover, and Russell Purvis. 2011. Technostress: technological antecedents and implications. *MIS quarterly* 35, 4 (2011), 831–858.
[9] Andrew Besmer and Heather Richter Lipford. 2010. Moving beyond untagging: photo privacy in a tagged world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1563–1572.
[10] YouTube Official Blog. 2012. Face blurring: when footage requires anonymity. *Blog.(18 July 2012). Retrieved April* 13 (2012), 2017.
[11] Alex Braunstein, Laura Granka, and Jessica Staddon. 2011. Indirect content privacy surveys: measuring privacy without asking about it. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*. 1–14.
[12] Kelly Erinn Caine. 2009. *Exploring everyday privacy behaviors and misclosures*. Ph. D. Dissertation. Georgia Institute of Technology.
[13] Kelly E Caine. 2009. Supporting privacy by preventing misclosure. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*. ACM, 3145–3148.
[14] Rich Caruana and Alexandru Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*. 161–168.
[15] Leucio Antonio Cutillo, Refik Molva, and Melek Önen. 2012. Privacy preserving picture sharing: Enforcing usage control in distributed on-line social networks. In *Proceedings of the Fifth Workshop on Social Network Systems*. ACM, 6.
[16] Maeve Duggan. 2013. Photo and video sharing grow online. *Pew Research Internet Project* (2013).
[17] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27, 8 (2006), 861–874.
[18] Andrea Frome, German Cheung, Ahmad Abdulkader, Marco Zennaro, Bo Wu, Alessandro Bissacco, Hartwig Adam, Hartmut Neven, and Luc Vincent. 2009. Large-scale privacy protection in google street view. In *2009 IEEE 12th international conference on computer vision*. IEEE, 2373–2380.
[19] Wenbo Guo, Dongliang Mu, Jun Xu, Purui Su, Gang Wang, and Xinyu Xing. 2018. LEMNA: Explaining Deep Learning based Security Applications. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 364–379.
[20] Rakibul Hasan, Eman Hassan, Yifang Li, Kelly Caine, David J Crandall, Roberto Hoyle, and Apu Kapadia. 2018. Viewer experience of obscuring scene elements in photos to enhance privacy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 47.
[21] Rakibul Hasan, Patrick Shaffer, David Crandall, Eman T Apu Kapadia, et al. 2017. Cartooning for enhanced privacy in lifelogging and streaming videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 29–38.
[22] Jianping He, Bin Liu, Deguang Kong, Xuan Bao, Na Wang, Hongxia Jin, and George Kesidis. 2016. Puppies: Transformation-supported personalized privacy preserving partial image sharing. In *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 359–370.
[23] Geoffrey E Hinton. 1999. Products of experts. (1999).
[24] Roberto Hoyle, Robert Templeman, Steven Armes, Denise Anthony, David Crandall, and Apu Kapadia. 2014. Privacy behaviors of lifeloggers using wearable cameras. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 571–582.
[25] Henry Hsu and Peter A Lachenbruch. 2014. Paired t test. *Wiley StatsRef: Statistics Reference Online* (2014).
[26] Hongxin Hu, Gail-Joon Ahn, and Jan Jorgensen. 2013. Multiparty access control for online social networks: model and mechanisms. *IEEE Transactions on Knowledge and Data Engineering* 25, 7 (2013), 1614–1627.
[27] Panagiotis Ilia, Iasonas Polakis, Elias Athanasopoulos, Federico Maggi, and Sotiris Ioannidis. 2015. Face/off: Preventing privacy leakage from photos in social networks. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 781–792.
[28] Katarzyna Janocha and Wojciech Marian Czarnecki. 2017. On loss functions for deep neural networks in classification. *arXiv preprint arXiv:1702.05659* (2017).
[29] Sanjay Kairam, Joseph'Jofish' Kaye, John Alexis Guerra-Gomez, and David A Shamma. 2016. Snap decisions?: How users, content, and aesthetics interact to shape photo sharing behaviors. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 113–124.
[30] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
[31] Thomas R Knapp. 1990. Treating ordinal scales as interval scales: an attempt to resolve the controversy. *Nursing research* 39, 2 (1990), 121–123.
[32] Balachander Krishnamurthy and Craig E Wills. 2009. On the leakage of personally identifiable information via online social networks. In *Proceedings of the 2nd ACM workshop on Online social networks*. ACM, 7–12.
[33] Yifang Li, Nishant Vishwamitra, Bart P Knijnenburg, Hongxin Hu, and Kelly Caine. 2017. Effectiveness and users' experience of obfuscation as a privacy-enhancing technology for sharing photos. *Manuscript submitted for publication* 4 (2017).
[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
[35] Erika McCallister, Timothy Grance, and Karen A Scarfone. 2010. Guide to protecting the confidentiality of personally identifiable information (PII). *Special Publication (NIST SP)-800-122* (2010).
[36] Andrew D Miller and W Keith Edwards. 2007. Give and take: a study of consumer photo-sharing culture and practice. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. Acm, 347–356.
[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
[38] Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2017. Searching for activation functions. *arXiv preprint arXiv:1710.05941* (2017).
[39] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.
[40] Anna Cinzia Squicciarini, Mohamed Shehab, and Federica Paci. 2009. Collective privacy management in social networks. In *Proceedings of the 18th international conference on World wide web*. ACM, 521–530.
[41] Anna Cinzia Squicciarini, Smitha Sundareswaran, Dan Lin, and Josh Wede. 2011. A3p: adaptive policy prediction for shared images over popular content sharing sites. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*. ACM, 261–270.
[42] Kimia Tajik, Akshith Gunasekaran, Rhea Dutta, Brandon Ellis, Rakesh B Bobba, Mike Rosulek, Charles V Wright, and Wu-chi Feng. 2019. Balancing Image Privacy and Usability with Thumbnail-Preserving Encryption. *IACR Cryptology ePrint Archive* 2019 (2019), 295.
[43] Shan-Tair Wang, Mei-Lin Yu, Chi-Jen Wang, and Chao-Ching Huang. 1999. Bridging the gap between the pros and cons in treating ordinal scales as interval scales from an analysis point of view. *Nursing research* 48, 4 (1999), 226–229.
[44] Mike Wu and Noah Goodman. 2018. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*. 5575–5585.
[45] Jun Yu, Baopeng Zhang, Zhengzhong Kuang, Dan Lin, and Jianping Fan. 2017. iPrivacy: image privacy protection by identifying sensitive objects via deep multi-task learning. *IEEE Transactions on Information Forensics and Security* 12, 5 (2017), 1005–1016.
[46] S Zerr, JH Stefan Siersdorfer, and E Demidova. 2012. Picalert! data set.
[47] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.