PrivObfNet: A Weakly Supervised Semantic Segmentation Model for Data Protection

ChiatPin Tay
Institue of High Performance Computing
A*STAR, Singapore
taycp@ihpc.a-star.edu.sg

Vigneshwaran Subbaraju
Institue of High Performance Computing
A*STAR, Singapore
vsubbaraju@ihpc.a-star.edu.sg

Thivya Kandappu
School of Computing and Information Systems
Singapore Management University, Singapore
thivyak@smu.edu.sg

Abstract

The use of social media has made it easy to communicate and share information over the internet. However, it also brings issues such as data privacy leakage, which can be exploited by recipients with malicious intentions to harm the sender. In this paper, we propose a deep neural network that analyzes user's image for privacy sensitive content and automatically locates sensitive regions for obfuscation. Our approach relies solely on image level annotations and learns to (a) predict an overall privacy score, (b) detect sensitive attributes and (c) demarcate the sensitive regions for obfuscation, in a given input image. We validated the performance of our proposed method on three large datasets, VISPR, PASCAL VOC 2012 and MS COCO 2014, in terms of privacy score, attribute prediction and obfuscation performance. On the VISPR dataset, we achieved a Pearson correlation of 0.88 and a Spearman correlation of 0.86, outperforming previous methods. On PASCAL VOC 2012 and MS COCO 2014, our model achieved a mean IOU of 71.5% and 43.9% respectively, and is among the state-of-the-art techniques using weakly supervised semantic segmentation learning.

1. Introduction

As of October 2022, more than half of the world's population uses social media for various purposes, including sharing personal information, marketing products, education, entertainment, and social activism, among others [31]. With the widespread adoption of smartphones and the rapid installation of transmission infrastructure, such as WiFi and 5G networks, a vast amount of data is shared at any given time. This data comes various forms, including text, images, audio, or videos. However, senders may not realize that they are sharing personal information when they post pictures of their bank cards, reveal their home address, or discuss their medical conditions online. As a result, this could lead to negative consequences such as professional reputation damage, identity theft, or financial fraud. These issues motivated many works to be done to protect data privacy [5, 11, 51, 57].

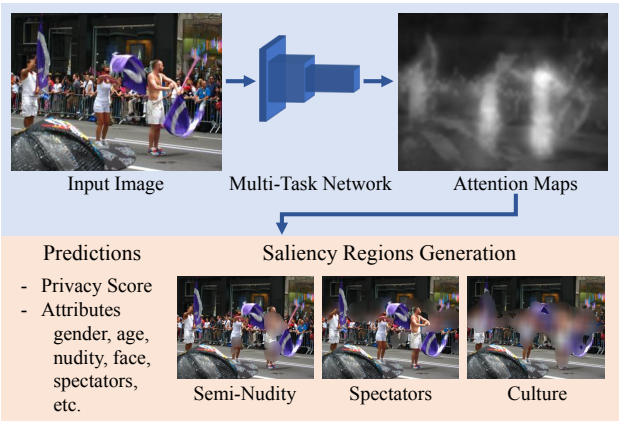

Figure 1. An overview of our proposed PrivObfNet. Attention maps are generated from the network and used for privacy score and attribute learning. Simultaneously, weakly supervised semantic segmentation learning is utilized to generate attribute-related saliency regions for obfuscation. Best viewed in color.

While it is ultimately the responsibility of the owner of personal information or data to carefully review content before sending, we believe that a more effective solution to prevent accidental sharing of data is to utilize an intelli-

gent system to analyze content and obfuscate sensitive areas before sharing on social media platforms. This approach ensures that users retain control over what information is shared and what is kept private, and also provides a personalized level of protection against accidental data sharing. Obfuscation decisions should be based on the user's preferences and cover both graphic and text content. This task is suitable for a deep learning computer vision model that can predict person attributes [48,61,63], compute privacy scores [11, 51, 71, 73], generate semantic segmentation or saliency regions [14, 22, 56, 75], and classify objects [13, 15, 54] with a high level of confidence. However, training such models typically requires large amount of data to correctly detect and segment the sensitive regions. Obtaining such training data is challenging, especially with the need of pixel-level semantic segmentation ground truth, as it is costly and time-consuming to prepare such datasets. In short, we require a deep learning model that can handle multiple privacy detection tasks with the constraint of learning from cheap labels only.

In this paper, we propose the PrivObfNet (Privacy Obfuscation Network) to automatically obfuscate a user's image based on their preferred sensitive settings. The proposed model is a multi-task network, consisting of three main tasks. The first task predicts the privacy score of an image, allowing the user to assess the sensitivity of their images and to decide whether to share them or not. The second task predicts privacy attributes such as gender, nudity, name, face, age, profession, culture, etc., allowing the user to prevent any sharing of data if sensitive attributes are detected in the image. The final task of PrivObfNet generates a mask for each attribute, enabling sensitive areas to be obfuscated while leaving the rest of the image untouched. An example of the workflow of our model is shown in Figure 1. In this example, the user chooses to hide three private content - semi-nudity, spectators and culture. Masks for the respective attributes are produced by the network and used to obfuscate the sensitive areas. Simultaneously, privacy score and presence of private attribute are predicted for the user to decide the next course of action.

The main challenge for our proposed model, and in fact for any semantic segmentation model, is the lack of pixel-level annotation for the class masks of most customized datasets. To address this issue, the computer vision community has explored various training methods, such as unsupervised learning [9, 12, 27, 52], self-supervised learning [3, 18, 25, 65] or weakly-supervised semantic segmentation (WSSS) learning method [2, 8, 28, 36, 37, 49, 53, 62, 66, 69, 70]. Our proposed approach adopts the WSSS method, utilizing class labels for object detection and semantic segmentation generation. This approach is cost-effective and can be easily implemented by most researchers and developers. Inspired by prior works [10, 28, 61, 65], we incorporate both global and local feature extraction in our model to improve the discriminative level of the feature vectors, thereby enhancing the generation of the obfuscation masks. We also integrated the privacy score and attribute tasks into our model, so that the user can decide the privacy level of their image, which is based on their personal preferences.

We benchmark the performance of our method using the VISPR data privacy dataset [51], the PASCAL VOC 2012 dataset [17] and the MS COCO 2014 dataset [46]. On the VISPR dataset, we achieve a Pearson correlation coefficient of 0.88 and a Spearman correlation coefficient of 0.86 for the privacy score, outperforming previous methods. On the PASCAL VOC 2012 and the MS COCO 2014 dataset, we achieved a mean Intersection-Over-Union (IOU) accuracy of 71.5% and 43.9% respectively, placing our method among the state-of-the-art techniques.

Our key contributions are:

1. We propose a weakly-supervised ensemble semantic segmentation architecture that encourages variations in the learning stage, thereby allowing the models to generate better masks for the data privacy obfuscation task.

2. We propose a multi-task privacy score prediction model that utilizes global features, local features, attribute learning, and privacy score teaching.

2. Related Works

We propose PrivObfNet, a deep learning model trained using weakly supervised semantic segmentation, to automatically obfuscate graphical or text content in an image. In this section, we describe various techniques developed over the years.

2.1. Personal Data Sharing and Score Prediction

There have been several studies [4, 6, 20, 38] examining what personal information is shared on social media and why. Commonly shared private information online includes name, gender, photo, birthday, education background, nationality, etc., and is usually accessed by friends [6]. Other studies [21, 30, 72] have investigated ways to mitigate accidental sharing of personal data. [51] conducted experiments to assess the consistency of users' privacy preferences, and proposed using a deep learning model to predict privacy score reliably after knowing users change their privacy preferences from time to time from survey respondents. They also provided the VISPR dataset, which contains 22,000 images annotated with 68 attributes, which we used for the privacy score training and benchmarking. [11] proposed a LSTM model to predict privacy scores and obfuscate the sensitive areas using attention maps. While predicting privacy scores is a useful way to inform users about the sensitivity of their data, the model is based on pre-assigned score

values derived from survey responses from 305 workers and may not be applicable to users with differing preferences. Thus, a more generic solution, such as private attribute prediction, should be provided.

2.2. Weakly Supervised Semantic Segmentation

Training the deep learning models without expensive labels has been the focus of many researchers. Techniques such as unsupervised learning [9, 12, 27, 52], self-supervised learning [3, 18, 25, 65] or weakly-supervised learning [8, 28, 49, 53, 62, 66, 69] are making great progress in recent years, each with their own strengths and weaknesses. Generally, strategies like unsupervised learning or self-supervised learning requires huge amount of unlabeled training data in order for the learning model to perform near at the level of their supervised learning counterpart. For studies relying on smaller dataset, a better approach is to leverage the image level annotation to guide the learning, and that's why WSSS has been an important research topic for the semantic segmentation domain.

Many WSSS approaches and techniques are proposed in recent years. For example, a popular approach known as attention map-based method attempts to generate saliency region from the feature maps. However, such method [76] relies heavily on the model's classifier, leading to only the hottest regions being identified as salient regions. [29] proposed a recursive method that can identify more object areas and simultaneously reduce out-of-boundary regions. Additionally, they introduced EdgePredictMix, a data augmentation method that calculates the probability difference information between adjacent pixels to define better object edges. Another recent work by [69] proposed the use of transformer models to generate the attention map. Their Multi-class Token Transformer (MCTformer) model can produce class-discriminative object localization maps by learning the interaction between multiple class tokens and the patch tokens. Furthermore, the location maps can be enhanced using the patch-level pairwise affinity coming from the patch-to-patch attention maps. These improvements in attention map generation have shown promising results in various WSSS tasks.

2.3. Attribute Learning

The process of learning person attributes has been demonstrated in several works [39, 41, 43, 47, 50, 64]. It involves extracting low-level features from input image to predict attributes. This technique is typically deployed in a multi-task system, where main task and attribute task are executed simultaneously, influencing each others. Since differing representations are coming from the multiple tasks, the model is exposed to more data variations, thus producing more discriminative descriptors. Additional information, such as attribute saliency regions, can also be extracted from the feature maps coming from the architecture backbone, allowing the user to localize the sensitive areas that represent the attribute. The ability to predict attributes allows the user to set privacy preferences after the model is trained, making it usable by any users, overcoming the constraints posed by the privacy score prediction approach where pre-determined privacy scores are needed as ground truth.

In the work [47], the global and attribute networks are learned jointly, with attribute score re-weighted to account for dependencies and correlations between attributes. For example, certain items such as skirt and handbag are more likely to be associated with female pedestrian rather than male pedestrian. To address this, the authors proposed an Attribute Re-weighting Module that transforms the predicted attribute score using trainable parameters, resulting in a new score that incorporates this correlation. Finally, the descriptor is obtained by concatenating the new score and the global feature vector.

The work by [50] introduced a Graph Convolutional Network (GCN) that depicts the dependencies between person attributes and local part features by processing the their correlation. This approach is similar to the one proposed by [47], where a re-weighting module is used to dynamically learn the correlations among attributes. However, in the GCN approach, a graph is constructed using the correlation between the attributes and the local part features, which enables the GCN to produce more robust representations for attribute prediction. To generate the human part masks required for the GCN, an off-the-shelf human parsing model called Self-Correction for Human Parsing (SCHP) [40] is used. SCHP is pre-trained on the LIP [44] dataset, and the generated masks are used as input to the GCN.

3. The Proposed Method

The proposed Privacy Obfuscation Network (PrivObfNet) is a multi-task model-ensemble deep learning framework designed for data privacy protection, and is shown in Figure 2. It can generate masks for obfuscation of sensitive areas in an image and also predict its privacy score and the privacy attributes. This allows individual user to decide what privacy setting to use for data protection. The PrivObfNet consists of two models: one for the global feature, and the other for the cropped image features. Each model processes attribute, privacy and global/cropped features, and generates attribute attention maps. Off-the-shelf pre-trained human parsing and OCR models are used to generate saliency masks of the training data, and are used to refine the object saliency regions of the training data. The framework computes four losses - global, cropped, attribute and privacy losses. Detailed descriptions of our method can be found below.

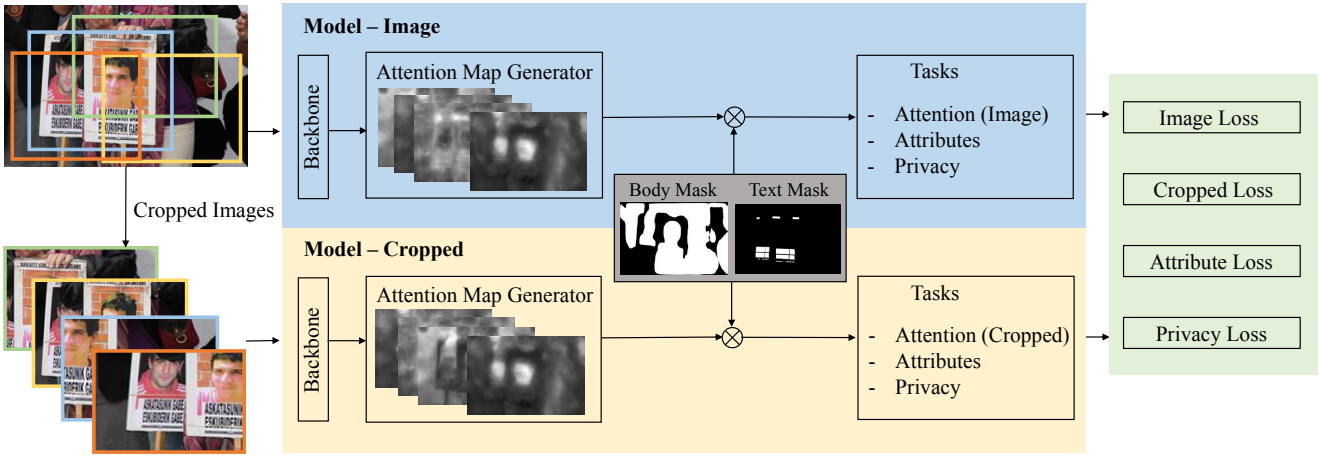

Figure 2. The PrivObfNet is a multi-task model-ensemble network consists of two models - Model-Image and Model-Cropped. Model-Image works on image level features, while Model-Cropped extracts local level features. Both models include an Attention Map Generator, and the Attention, Attribute and Privacy tasks. The attention maps are weighted by the body and the text masks if necessary, and are used for further processing by the subsequent tasks to generate feature vectors for loss calculation. The knowledge acquired by the Model-Cropped are transferred to the Model-Image via the loss functions, allowing higher attribute and privacy score accuracies, and better semantic masks prediction for obfuscation.

3.1. Multi-task Model-Ensemble Framework

Model ensemble [74] is a well-established technique for improving the prediction capabilities of models. In our study, we propose a multi-task network using two similar models. The first model, Model-Image, is for learning image-level features, which are typically used in many baseline computer vision task, such as object classification or person re-identification. The second model, Model-Cropped, is designed to learn finer or localized area of the image, which allows the network neurons to activate salient regions normally ignored at global level. The input image is randomly cropped into a few smaller images and then passed through the backbone and CNN layers to produce the necessary feature maps for attention map generation and loss calculation. The overall framework is illustrated in Figure 2. The backbone can be any standard architecture such as ResNet [24], VGG [58] or Inception [60]. We chose ResNet-38 [68] as it is commonly used by WSSS community and has good performance over other backbones.

The PrivObfNet consists of three main tasks: Attention, Privacy and Attribute. These tasks create attention maps, image features (Model-Image) or cropped features (Model-Cropped), privacy features and attribute features. Four losses are computed from these features. During inference, the Model-Image, which learns using complete input image, is used to predict the privacy score and attributes and to generate attribute masks for obfuscation.

We like to present the total loss now, and then elaborate on the individual losses in subsequent sections. The total loss can be described as:

$$\mathcal{L}_{all} = \lambda_1 \mathcal{L}_{img} + \lambda_2 \mathcal{L}_{cr} + \lambda_3 \mathcal{L}_{priv} + \lambda_4 \mathcal{L}_{att} \quad (1)$$

where $\mathcal{L}_{all}$, $\mathcal{L}_{img}$, $\mathcal{L}_{cr}$, $\mathcal{L}_{priv}$ and $\mathcal{L}_{att}$ denote total loss, image loss, cropped loss, privacy loss and attribute loss respectively. λ_i is the weighing factor for each of the losses.

3.2. Attention Map Generator

The Attention Map Generator. This module computes the attribute attention maps from the feature maps and send them to the subsequent Attention, Attribute and Privacy tasks to produce the respective feature vectors for loss calculation. We adopted the Class Activation Mapping (CAM) [75] technique to produce the initial heatmaps for both models. To discover as much saliency regions belonging to the class, we randomly cropped the input image into multiple smaller ones and fed them to the Model-Cropped. This forces the Model-Cropped to produce additional desired saliency regions not found in the Model-Image. With our model ensemble setup, the final learned saliency mask is a combination of all the discovered heatmaps, producing a much accurate representation of the privacy attribute shape.

Human and Text Masks. The main objective of WSSS is to achieve good semantic segmentation training with the lack of pixel-wise class labeling as it is costly to annotate them. To predict a good mask that has well defined boundary, large coverage of the object and removal of the background, we can leverage off-the-shelf human parsing [19, 23, 45, 54] and text recognition [16, 26] models that are trained on large and fully annotated dataset to generate

clean masks in advance and use them to guide our framework. As shown in Figure 2, the generated human mask and text mask are used to weigh the attribute attention maps from both models, removing background noise for each attribute maps. Table 1 shows examples of attributes and their corresponding mask types. Based on how well the attributes are found in the image, they are categorize into three mask types - Un-Masked, Human Mask and Text Mask. For example, Age and Full Name are assigned Human and Text masks respectively because their corresponding image regions can be extracted accurately. Student ID or Legal Involvement attribute, however, require more information than just human and text masks, thus it is better to use whole attention maps (Un-Masked) for learning. All attribute attention maps are then concatenated along the channel dimension to form the final attention map.

3.3. Attention Task

The Attention Task consists of the Image and Cropped Losses. Image Loss works on global image feature whereas Cropped Loss uses cropped image features for processing.

Image Loss. Given an image I, we first generate the global saliency map G^c of class c from Model-Image. Next, we generate cropped saliency maps $\{S_1^c, S_2^c, ..., S_N^c\}$ for Model-Cropped, where N is the number of cropped images. Since the global saliency map G^c is spatially larger then the cropped saliency maps S_i^c, we crop G^c to produce $\{G_1^c, G_2^c, ..., G_N^c\}$ so that both sets of saliency maps have matching size and spatial position. In another word, both G_i^c and S_i^c are cropped at the same positions when referred back to the input image. The G_i^c is normalized using a Softmax function along the channel direction. We can now form N pairs of saliency maps, i.e. (G_1^c, S_1^c) and use them to calculate the image loss, as follows

$$\mathcal{L}_{img} = \frac{1}{N}\sum_{i=1}^{N} ||Softmax(G_i^c) - S_i^c||^2 \qquad (2)$$

Cropped Loss. This is a regression loss that utilizes only the attention maps S_i from Model-Cropped. First, the attention maps are transformed into feature vectors through global average pooling. Then, the multi-label soft margin loss is computed as shown in Equation 3. Since the object classifier tends to amplify class's hot spots, this loss helps to boost the saliency regions found in the cropped images. It also improves the prediction accuracy of attribute task and privacy score task.

$$\mathcal{L}_{cr} = \frac{-1}{NC}\sum_{i=1}^{N}\sum_{c=1}^{C}\begin{cases} y^c log(\frac{1}{1+exp(-f_i^c)}) & \text{if } y^c = 1 \\ (1-y^c) log(\frac{exp(-f_i^c)}{1+exp(-f_i^c)}) & \text{if } y^c = 0 \end{cases} \qquad (3)$$

Un-Masked	Mask Type Human	Text
Safe	Age	Full Name
Credit Card	Weight	Birth City
Student ID	Eye Color	Hand Writing
Ticket	Nudity	Maritus Status
Legal Involvement	Race	Birth Date

Table 1. Examples of mask type for image attributes. The attributes of an image can be assigned Un-Masked, Human or Text mask types, depending on how well they are represented by the mask.

where y^c is the label for class c, C is the number of classes and f_i^c is the feature vector of class c and cropped image i

3.4. Privacy Task

Before sharing any image on the social media, most users will assess if there is any private information in it. However, the experiment conducted by [51] showed that users may not be consistent in their privacy judgements and make conflicting decisions when sharing data online. Thus, it would be beneficial if a deep learning model, which is capable of predicting the privacy score of an image with high confidence, can alert the user of the level of potential risk involved. However, such approach requires the user's privacy preferences to be available during the training time, thereby making the model user-specific and not usable by others. Nevertheless, we believe that privacy score prediction is still useful for a group of users having similar preferences. Additionally, it can be a good indicator of our model's prediction accuracy. The privacy loss function is as follows,

$$\mathcal{L}_{pri} = \frac{1}{N}\sum_{i=1}^{N} MSE(f(V), y_i) \qquad (4)$$

where $f(V)$ is the privacy feature vector, y_i is the user privacy preference, and $MSE()$ is the mean squared error function. Note that the privacy preferences are provided on the VISPR dataset, but not on PASCAL VOC 2012 and MS COCO 2014. Thus we only incorporate this task when benchmarking using the VISPR dataste.

3.5. Attribute Task

The attribute task enables us to predict the presence or absence of personal sensitive attributes like gender, name, face or nudity within the input image. Each of the attributes is individually trained using cross-entropy loss, which is defined in Equation 5.

$$\mathcal{L}_{att} = -\frac{1}{NC}\sum_{i=1}^{N}\sum_{c=1}^{C} y^c log(q_i^c) + (1-y^c) log(1-q_i^c) \quad (5)$$

where y^c is the label for class c and q_i^c is the softmax output of the feature vector belong to class c and cropped image i. This task uses only the feature maps from Model-Image only.

4. Experiments

We conducted experiments to characterize and validate the performance of our proposed PrivObfNet using the VISPR [51], PASCAL VOC2012 [17] and MS COCO 2014 dataset [46]. The three important accuracy tests are privacy score prediction, privacy attribute prediction and attention map generation. The first two tests are conducted using VISPR dataset, and the last test is verified qualitatively using VISPR dataset but quantitatively using PASCAL VOC 2012 and MS COCO 2014 dataset.

VISPR dataset consists of 10,000 training images, 4,167 validation images and 8,000 test images. These images are collected mainly from the OpenImages dataset [33], with some images from Twitter to reduce the imbalance of a few attributes like credit cards. There is a total of 68 privacy attributes, which include gender, age, name, weight, handwriting, occupation, address, and more. On average, there are 761 images per attribute. However, text-related attributes have lower number of images in the dataset. Those attributes with less than 50 images include Legal Involvement, Nudity, Fingerprint and Home Address.

Both the PASCAL VOC 2012 and MS COCO 2014 are widely used computer vision dataset for object detection, segmentation and classification tasks. They are annotated with image-level class labels, bounding boxes and pixel-level semantic segmentation labeling, and with 21 and 81 classes respectively.

4.1. Implementation

We chose Resnet-38 as our backbone for both Model-Image and Model-Cropped as it is commonly used by the WSSS community, thereby allowing us to have a fair comparison with previous published methods. Both models are initialized with pretrained ImageNet's weights [13]. The input image size is 448x448 for the Model-Image and 320x320 for the Model-Cropped. They were chosen to balance computational efficiency and model performance. The framework was trained for 10 epoch, with decay at the sixth epoch, and the learning rate was set to 0.002 with a weight decay of 5e-04. SGD optimizer was used for training, and the batch size and patch size were set to 4 and 6 respectively.

The loss weights were set as follows: $\lambda_1 = 10$ for the image loss, $\lambda_2 = 1$ for the cropped loss, $\lambda_3 = 1$ for the privacy score loss and $\lambda_4 = 1$ for the privacy attribute loss. The loss weights were chosen to balance the contributions of the different components of the model during training.

Methods	L_1-Error	Correlation ρ_p	Correlation ρ_s
AP-PR [51]	0.66	-	-
PR-CNN [51]	0.64	-	-
PrivAttNet$_{MLC}$ [11]	0.44	0.83	0.76
PrivNet [11]	0.43	0.83	0.78
PrivAttNet [11]	0.40	0.87	0.84
PrivObfNet (Ours)	**0.09**	**0.88**	**0.86**

Table 2. Privacy score performance comparison with other published methods on VISPR dataset.

4.2. VISPR Dataset

In this section, we discuss all the experiments conducted using the VISPR dataset, including privacy score prediction, obfuscation visualization and attribute prediction.

4.2.1 Privacy Score

The results for the privacy score are shown in Table 2. We used two metrics in our experiments, which are also used by [11, 51] to evaluate the scores. The first metric, L_1-Error, is computed by averaging the absolute differences between the predicted scores and the labels. The second metric uses both Pearson and Spearman correlation coefficients for benchmarking. As shown, PrivObfNet has the lowest L_1-Error, and the highest Pearson and Spearman correlation coefficients, indicating our framework provides the best match between predicted scores and the user preferences.

4.2.2 Visualization of Privacy Obfuscation

Using model trained on VISPR dataset, we provide examples of the images with their privacy attribute obfuscated in Figure 3 for qualitative visualization. As shown, attributes related to human, such as face, race or occupation, have masks in finer granularity than text content-related attributes. The generated masks for Address, Handwriting, Opinion, Online Conversation or Name attributes, which are all text related, tend to cover most of the text contents within the image. There are two reasons for this. Firstly, many text-related attributes have insufficient training images, as explained in the introductory paragraphs of Section 4 Experiment. Secondly, the text mask used for refining the attention map covers all textual areas, and not specific localized attributes. Thus producing precise masks to obfuscate exact privacy text attribute is challenging. This problem can

Figure 3. Visualization of privacy obfuscation on VISPR dataset images. The generated masks are effective in obfuscating the privacy areas in the images. Note that as the dataset contains more images with human subjects than those with text content, the attention masks generated for human-related attributes (face, race, or injury) exhibit finer granularity as compared to those for text-related attributes (address, handwriting or ticket).

be solved if sufficient images are provided for training, as illustrated by the obfuscation of human attributes.

4.2.3 Privacy Attribute Prediction

The accuracy of attribute prediction is a crucial factor in evaluating the effectiveness of our framework as a generic privacy obfuscation tool. In Table 3, we present the accuracy results obtained using the Attribute Task network described in Section 3.5. Our results demonstrate a strong prediction accuracy from PrivObfNet, achieving a mean 92.8% for attributes with more than 300 training images. The choice of the number of training samples per class is based on the ratio of 1:10 (1/10 of the maximum number of training images per class) to ensure reliable and balanced predictions. Attributes such as License Plate, Email Content, Username. Religion, Birth City or Ethnic Clothing were excluded due to the lack of training samples.

Privacy Attribute Accuracy (%)			
Attribute	Accuracy	Attribute	Accuracy
Safe	88.1	Age	91.7
Weight	85.1	Height	85.1
Gender	93.6	Eye	87.3
Hair	93.1	Face	90.4
Face	87.2	Semi Nude	96.2
Race	93.4	Color	94.2
Full Name	94.6	Occupation	92.5
Occasion	96.8	Culture	95.7
Sports	97.3	Opinion	97.1
Personal	96.9	Social	92.9
Professional	94.7	Spectator	93.2
Viewers	97.1	Landmark	91.3
Address	92.7	Date Time	93.7

Table 3. Privacy attribute accuracy performance on VISPR dataset. To mitigate the issue of data imbalance, only attributes with more than 300 training images are shown here. The mean accuracy is 92.8%.

4.2.4 Ablation Study

We performed an ablation study to investigate the impact of multitasking on the performance of our PrivObfNet. Table 4 summarizes the results obtained from different task configurations. Our baseline model consists only of the Attention (Image) task, and it achieved Pearson and Spearman correlation coefficients of 0.79 and 0.76, respectively. Adding the Privacy task improved the coefficients to 0.82 and 0.79, while replacing the Privacy task with the Attribute task resulted in coefficients of 0.82 and 0.80. Combining both Privacy and Attribute tasks with the baseline further improved the coefficients to 0.85 and 0.83, respectively. Finally, we incorporated Cropped loss to the network and the coefficients were pushed up to 0.88 and 0.86, respectively. These

Tasks	Correlation ρ_p	ρ_s
Attention (Image)	0.79	0.76
Attention (Image) + Priv	0.82	0.79
Attention (Image) + Attr	0.82	0.80
Attention (Image) + Priv + Attr	0.85	0.83
Attention (Image + Cropped) + Priv + Attr	0.88	0.86

Table 4. Ablation study of PrivObfNet using multi-task approach. Priv and Attr denote Privacy and Attribute respectively.

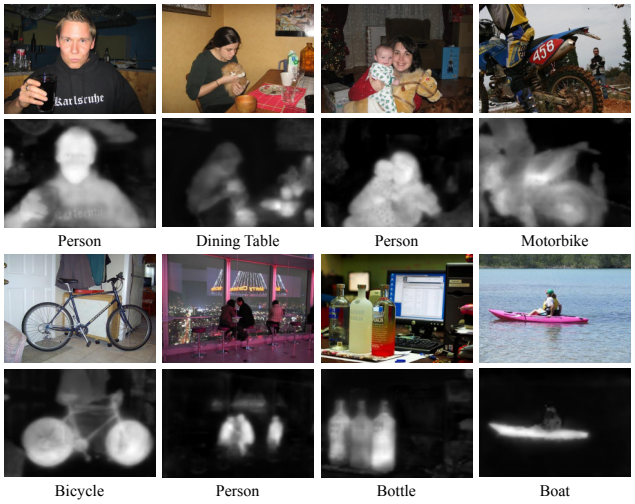

Figure 4. Visualization of generated attention maps on PASCAL VOC 2012 dataset.

findings suggest that concurrent learning of relevant tasks can lead to a synergistic effect, resulting in more discriminative descriptors and better prediction of privacy score and attributes compared to prior methods.

4.3. PASCAL VOC 2012 and MS COCO 2014

We use both PASCAL VOC 2012 and MS COCO 2014 dataset to quantitatively validate the semantic segmentation performance of our proposed PrivObfNet. Following the existing WSSS research studies [35, 42, 55, 67], we adopted the typical evaluation metric for both dataset, and we used the DeepLabV2 ResNet101 [7] as the backbone CNN. Both dataset are trained and benchmark with the state-of-the-art using the mIOU. As shown in Table 5, our multi-task approach yielded excellent result of 71.5% and 43.9% on PASCAL VOC 2012 and MS COCO dataset respectively. Figure 4 depicts the generated attention maps, produced by our model, which provide excellent qualitative insight into the PrivObfNet semantic segmentation output.

Methods	Sup.	VOC test	COCO val
EDAM (cvpr'21) [67]	I+S	70.6	-
EPS++ (PAMI'23) [35]	I+S	72.4	42.4
DRS (AAAI'21) [32]	I+S	71.4	-
L2G (cvpr'22) [28]	I+S	71.7	44.2
RCA (cvpr'22) [77]	I+S	**72.8**	36.8
IRN (cvpr'19) [1]	I	64.8	41.4
CDA (ICCV'21) [59]	I	66.8	33.2
RIB (NeurIPS'21) [34]	I	68.6	43.8
URN (AAAI'22) [42]	I	69.7	40.7
MCTformer (cvpr'22) [69]	I	71.6	42.0
BECO (cvpr'23) [55]	I	71.8	**45.1**
PrivObfNet (Ours)	I+S	71.5	43.9

Table 5. Comparison of pseudo segmentation labels with state-of-the-art on PASCAL VOC 2012 test set and MS COCO val set. I and S denote image-level labels and pre-generated saliency masks respectively.

5. Conclusion

This paper presents a novel multi-task model-ensemble WSSS deep learning framework for data privacy protection. The model consists of three main tasks - Attention, Attribute and Privacy tasks. The proposed framework utilizes four loss functions - image, cropped, attribute and privacy losses - to guide the learning process. During inference, the framework generates attention maps, predicts privacy attributes, and computes privacy scores to assist users in their privacy data protection efforts. Attention maps of privacy attributes such as gender, age, name, birth city and nudity, among others, are generated and used to obfuscate images. Users can evaluate the risk of sharing images using the privacy attributes prediction or privacy score.

We validated the performance of our approach using VISPR, PASCAL VOC 2012 and MS COCO 2014 dataset. On VISPR dataset, we achieved Pearson and Spearman correlation coefficients of 0.88 and 0.86, respectively, outperforming prior methods. On the PASCAL VOC and COCO dataset, we attained a mIOU of 71.5% and 43.9% respectively using the DeeplabV2 Resnet101 model, placing our method among the state-of-the-art.

6. Acknowledgment

This research is supported by Agency for Science, Technology and Research (A*STAR) project No. 202D800021.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2209–2218, 2019. 8

[2] Peri Akiva and Kristin Dana. Single stage weakly supervised semantic segmentation of complex scenes. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5943–5954, 2023. 2

[3] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15384–15394, 2021. 2, 3

[4] Iqra Bashir, Amara Malik, and Khalid Mahmood. Social media use and information-sharing behaviour of university students. *IFLA journal*, 47(4):481–492, 2021. 2

[5] Tânia Carvalho, Nuno Moniz, Pedro Faria, and Luís Antunes. Towards a data privacy-predictive performance trade-off. *Expert Systems with Applications*, page 119785, 2023. 1

[6] Chi Kin Chan and Johanna Virkki. Perspectives for sharing personal information on online social networks. *Social Networking*, 2014, 2014. 2

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. 8

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 2, 3

[9] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. *Advances in neural information processing systems*, 32, 2019. 2, 3

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[11] Zhang Chen, Thivya Kandappu, and Vigneshwaran Subbaraju. Privattnet: Predicting privacy risks in images using visual attention. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10327–10334. IEEE, 2021. 1, 2, 6

[12] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16794–16804, 2021. 2, 3

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 6

[14] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Semantic correlation promoted shape-variant context for segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8885–8894, 2019. 2

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. *arXiv preprint arXiv:2010.11929*, 2010. 2

[16] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*, 2020. 4

[17] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. 2, 6

[18] Yann Fabel, Bijan Nouri, Stefan Wilbert, Niklas Blum, Rudolph Triebel, Marcel Hasenbalg, Pascal Kuhn, Luis F Zarzalejo, and Robert Pitz-Paal. Applying self-supervised learning for semantic cloud segmentation of all-sky images. *Atmospheric Measurement Techniques*, 15(3):797–809, 2022. 2, 3

[19] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 932–940, 2017. 4

[20] Ralph Gross and Alessandro Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80, 2005. 2

[21] Saikat Guha, Kevin Tang, and Paul Francis. Noyb: Privacy in online social networks. In *Proceedings of the first workshop on Online social networks*, pages 49–54, 2008. 2

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. 4

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[25] Lukas Hoyer, Dengxin Dai, Yuhua Chen, Adrian Koring, Suman Saha, and Luc Van Gool. Three ways to improve semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11130–11140, 2021. 2, 3

[26] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE, 2019. 4

[27] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification

and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019. 2, 3

[28] Peng-Tao Jiang, Yuqi Yang, Qibin Hou, and Yunchao Wei. L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16886–16896, 2022. 2, 3, 8

[29] Sanghyun Jo, In-Jae Yu, and Kyungsu Kim. Recurseed and edgepredictmix: Single-stage learning is sufficient for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2204.06754*, 2022. 3

[30] Imrul Kayes and Adriana Iamnitchi. Privacy and security in online social networks: A survey. *Online Social Networks and Media*, 3:1–21, 2017. 2

[31] Simon Kemp. Digital 2022: October global statshot report. *DataReportal*, 20 October 2022. , accessed 18 March 2023, ¡https://datareportal.com/reports/digital-2022-october-global-statshot>. 1

[32] Beomyoung Kim, Sangeun Han, and Junmo Kim. Discriminative region suppression for weakly-supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1754–1761, 2021. 8

[33] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017. 6

[34] Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon. Reducing information bottleneck for weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 34:27408–27421, 2021. 8

[35] Minhyun Lee, Seungho Lee, Jongwuk Lee, and Hyunjung Shim. Saliency as pseudo-pixel supervision for weakly and semi-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 8

[36] Min Seok Lee, Seok Woo Yang, and Sung Won Han. Gaia: Graphical information gain based attention network for weakly supervised point cloud semantic segmentation. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 582–591, 2023. 2

[37] Yuan-Hao Lee, Fu-En Yang, and Yu-Chiang Frank Wang. A pixel-level meta-learner for weakly supervised few-shot semantic segmentation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1607–1617, 2022. 2

[38] Amanda Lenhart and Mary Madden. How teens manage their online identities and personal information in the age of myspace washington dc: Pew internet & american life project. *Retrieved January*, 2008. 2

[39] Huafeng Li, Yiwen Chen, Dapeng Tao, Zhengtao Yu, and Guanqiu Qi. Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification. *IEEE Transactions on Information Forensics and Security*, 16:1480–1494, 2020. 3

[40] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 3

[41] Shuzhao Li, Huimin Yu, and Roland Hu. Attributes-aided part detection and refinement for person re-identification. *Pattern Recognition*, 97:107016, 2020. 3

[42] Yi Li, Yiqun Duan, Zhanghui Kuang, Yimin Chen, Wayne Zhang, and Xiaomeng Li. Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1447–1455, 2022. 8

[43] Yang Li, Huahu Xu, Minjie Bian, and Junsheng Xiao. Automatic attribute learning for person re-identification. In *Journal of Physics: Conference Series*, volume 1576, page 012007. IOP Publishing, 2020. 3

[44] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):871–885, 2018. 3

[45] Liang Lin, Yiming Gao, Ke Gong, Meng Wang, and Xiaodan Liang. Graphonomy: Universal image parsing via graph reasoning and transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2504–2518, 2020. 4

[46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 6

[47] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95:151–161, 2019. 3

[48] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang. Improving person re-identification by attribute and identity learning. *CoRR*, abs/1703.07220, 2017. 2

[49] Chunmeng Liu, Enze Xie, Wenjia Wang, Wenhai Wang, Guangyao Li, and Ping Luo. Wegformer: Transformers for weakly supervised semantic segmentation. *arXiv preprint arXiv:2203.08421*, 2022. 2, 3

[50] Binh X Nguyen, Binh D Nguyen, Tuong Do, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Graph-based person signature for person re-identifications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3492–3501, 2021. 3

[51] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE international conference on computer vision*, pages 3686–3695, 2017. 1, 2, 5, 6

[52] Yassine Ouali, Céline Hudelot, and Myriam Tami. Autoregressive unsupervised image segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 142–158. Springer, 2020. 2, 3

[53] Shun-Yi Pan, Cheng-You Lu, Shih-Po Lee, and Wen-Hsiao Peng. Weakly-supervised image semantic segmentation using graph convolutional networks. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 2, 3

[54] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2, 4

[55] Shenghai Rong, Bohai Tu, Zilei Wang, and Junjie Li. Boundary-enhanced co-training for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19574–19584, 2023. 8

[56] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, 2017. 2

[57] Paulo Silva, Carolina Gonçalves, Nuno Antunes, Marilia Curado, and Bogdan Walek. Privacy risk assessment and privacy-preserving data monitoring. *Expert Systems with Applications*, 200:116867, 2022. 1

[58] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[59] Yukun Su, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. Context decoupling augmentation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7004–7014, 2021. 8

[60] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 4

[61] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7134–7143, 2019. 2

[62] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7158–7166, 2017. 2, 3

[63] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Attribute recognition by joint recurrent learning of context and correlation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 531–540, 2017. 2

[64] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. *arXiv preprint arXiv:1803.09786*, 2018. 3

[65] Yuan Wang, Wei Zhuo, Yucong Li, Zhi Wang, Qi Ju, and Wenwu Zhu. Fully self-supervised learning for semantic segmentation. *arXiv preprint arXiv:2202.11981*, 2022. 2, 3

[66] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017. 2, 3

[67] Tong Wu, Junshi Huang, Guangyu Gao, Xiaoming Wei, Xiaolin Wei, Xuan Luo, and Chi Harold Liu. Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16765–16774, 2021. 8

[68] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019. 4

[69] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4310–4319, 2022. 2, 3, 8

[70] Xiaowen Ying, Xin Li, and Mooi Choo Chuah. Weakly-supervised object representation learning for few-shot semantic segmentation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1496–1505, 2021. 2

[71] Alyson L Young and Anabel Quan-Haase. Information revelation and internet privacy concerns on social network sites: a case study of facebook. In *Proceedings of the fourth international conference on Communities and technologies*, pages 265–274, 2009. 2

[72] Chi Zhang, Jinyuan Sun, Xiaoyan Zhu, and Yuguang Fang. Privacy and security for online social networks: challenges and opportunities. *IEEE Network*, 24(4):13–18, 2010. 2

[73] Guangsheng Zhang, Bo Liu, Tianqing Zhu, Andi Zhou, and Wanlei Zhou. Visual privacy attacks and defenses in deep learning: a survey. *Artificial Intelligence Review*, 55(6):4347–4401, 2022. 2

[74] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. *CoRR*, abs/1706.00384, 2017. 4

[75] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2, 4

[76] Sanping Zhou, Fei Wang, Zeyi Huang, and Jinjun Wang. Discriminative feature learning with consistent attention regularization for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8040–8049, 2019. 3

[77] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4299–4309, 2022. 8