ERASER: Adversarial Sensitive Element Remover for Image Privacy Preservation

Guang Yang[1,2,3], Juan Cao[2,3], Danding Wang[2], Peng Qi[2,3], Jintao Li[2]

[1] Zhongguancun Laboratory
[2] Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences
yangguang@zgclab.edu.cn, {caojuan, wangdanding, qipeng, jtli}@ict.ac.cn

Abstract

The daily practice of online image sharing enriches our lives, but also raises a severe issue of privacy leakage. To mitigate the privacy risks during image sharing, some researchers modify the sensitive elements in images with visual obfuscation methods including traditional ones like blurring and pixelating, as well as generative ones based on deep learning. However, images processed by such methods may be recovered or recognized by models, which cannot guarantee privacy. Further, traditional methods make the images very unnatural with low image quality. Although generative methods produce better images, most of them suffer from insufficiency in the frequency domain, which influences image quality. Therefore, we propose the Adv**ER**s**A**rial **S**ensitive **E**lement **R**emover (ERASER) to guarantee both image privacy and image quality. 1) To preserve image privacy, for the regions containing sensitive elements, ERASER guarantees enough difference after being modified in an adversarial way. Specifically, we take both the region and global content into consideration with a Prior Transformer and obtain the corresponding region prior and global prior. Based on the priors, ERASER is trained with an adversarial Difference Loss to make the content in the regions different. As a result, ERASER can reserve the main structure and change the texture of the target regions for image privacy preservation. 2) To guarantee the image quality, ERASER improves the frequency insufficiency of current generative methods. Specifically, the region prior and global prior are processed with Fast Fourier Convolution to capture characteristics and achieve consistency in both pixel and frequency domains. Quantitative analyses demonstrate that the proposed ERASER achieves a balance between image quality and image privacy preservation, while qualitative analyses demonstrate that ERASER indeed reduces the privacy risk from the visual perception aspect.

Introduction

People record and share their lives with a large number of images on social media platforms like Facebook and Instagram.

Sharing images is very convenient due to smartphones and mobile Internet, but such convenience brings the risk of privacy leakage at the same time. The shared images contain various types of sensitive information like income, disease, and home address (Orekondy, Schiele, and Fritz 2017), and indirect information can be recognized by data analysis models (Hu et al. 2018; Yang et al. 2018). The sensitive information leaked by images has been documented to be used maliciously (Solsman 2020) and caused severe consequences like fraud and cyber violence (Equifax 2020). As an urgent issue that is close to our daily life, image privacy preservation is attracting increasing concerns.

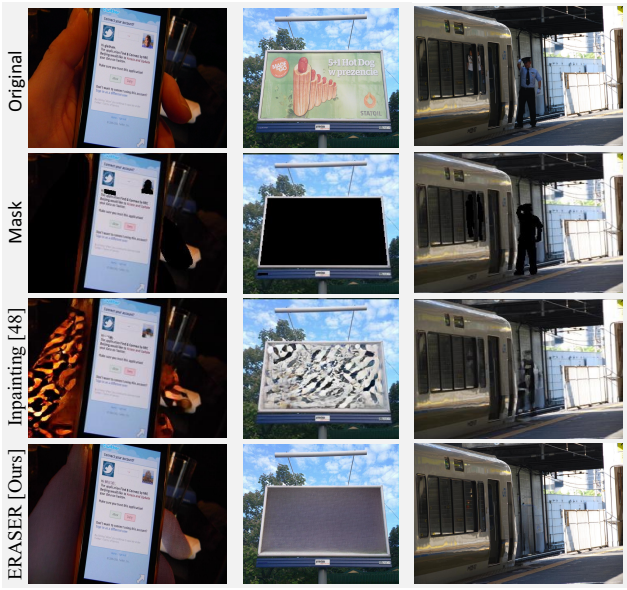

Figure 1: Samples of images (first row), sensitive elements masked by black blocks (second row), images edited with image inpainting (Wan et al. 2021) (third row), and images edited with ERASER (bottom row). (Zoom in for details)

Although social media platforms allow users to show their images only to specific people for privacy preservation, the images may be saved and forwarded. In addition, some platforms by default will analyze and recommend the image to people who may be interested. Such situations will make the content seen by undesirable people. Completely avoiding image sharing will be a safe choice but fail to meet the requirement of daily social life. Such phenomena and the potential harms make it urgent to design methods to achieve the balance between image sharing and privacy preservation.

An intuitive way is to modify the sensitive elements in the images to reduce the privacy risk. Some researchers adopt traditional image processing methods like pixelation (Fan 2018), blur, and mask (Li et al. 2017) to edit the sensitive elements. However, such methods cannot make the image looks natural after editing, which degrades social usability. More importantly, the editing traces are very obvious and malicious viewers could easily perceive that the images have been edited (Liu et al. 2020a). Further, the identity may be recognized by targeted detector (Oh et al. 2016), and the edited regions may be recovered (Song et al. 2018; Xiong et al. 2019; Shen et al. 2019). The traditional methods cannot well protect image privacy.

Recently, generation-based image inpainting that fills specific regions of the image and makes the entire image looks natural, seems to be a promising solution. For example, Uittenbogaard et al. (2019) proposed an inpainting framework to remove pedestrians and vehicles in street-view panoramas. However, several issues exist that only inpainting is not enough for image privacy preservation. First, the only task of image inpainting is to naturally fill the region to guarantee visual integrity. But for image privacy preservation, the original content of the regions should be considered to guarantee difference after edition. Second, there are often many small regions or several large regions that contain sensitive information. Most inpainting models take all regions as a whole and only consider the local context, which leads to unnatural results. Third, most current methods process the image only in the pixel domain, while some works have demonstrated that current generative methods are insufficient in the frequency domain which degrades the image quality (Jiang et al. 2021).

In this paper, we propose a new task of Sensitive Elements Removing (SER) to achieve a balance between image social usability and image privacy preservation. First, similar to image painting, SER needs to naturally edit the regions to guarantee the social usability of the images. Second, SER requires the content in the edited region different from the original one in two aspects, human perception and model recognition, to preserve image privacy. To this end, we propose an Adv**ER**s**A**rial **S**ensitive **E**lement **R**emover (ERASER), of which the workflow is presented in Fig. 2: (1) We take both the global and regional structures into consideration with Prior Transformer. We build the global prior P_G and region prior P_R with Transformer (Vaswani et al. 2017) for two reasons. First, Transformer can capture global information without the limitation of the receptive field. Second, the self-attention (i.e., the basic block of the Transformer) can model the correlation among the target regions. (2) To improve the insufficient performance in the frequency domain, the original image I and the prior P_G, P_R are processed with Fast Fourier Convolution (FFC) (Chi, Jiang, and Mu 2020) instead of the traditional convolution block. With FFC ERASER, the masked regions M are edited by considering both pixel and frequency domain characteristics to obtain the modified image $\hat{I}$. (3) Finally, to guarantee that the content in the edited regions has changed from the original one, a Difference Loss is combined with the general generative loss in an adversarial way to make the edited regions different.

A tremendous obstacle for SER is that there is no suitable dataset for this task. Image inpainting datasets can be obtained by adding noise to the image, but there are no ground truth image pairs of images with and without sensitive elements. To tackle this challenge, we designed a pipeline to start the training with the help of inpainting datasets, then utilize another dataset with the segmentation of sensitive elements for fine-tuning. Further, we design a reasonable pipeline to evaluate the proposed ERASER. Experimental results demonstrate that ERASER can achieve the balance of image usability and image privacy.

Our main contributions are summarized as follows:

(1) We propose the task, SER, for image privacy preservation. SER naturally fills specific regions and makes the modified regions different to achieve the balance of image quality and image privacy preservation.

(2) We propose an ERASER to solve the proposed task and design a reasonable pipeline to train and evaluate the method without enough image pairs of ground truth.

(3) To make the modified image more natural, ERASER is designed to capture global and region structures in both pixel and frequency domains. To make the edited regions different, the idea of the adversarial sample is combined with a well-designed training strategy.

(4) Experimental results demonstrate that ERASER achieved the goal in both quantitative and qualitative analyses. The edited image is natural, while the edited regions are different for both visual perception and model recognition.

Related Work

Image Privacy Preservation

The most direct privacy risk of images is the exposed sensitive information. Some researchers identify privacy-leaking images and prevent unintentional sharing (Zerr et al. 2012; Tonge and Caragea 2019; Yang et al. 2020, 2022). Such methods protect the image privacy well, but cannot meet the sharing requirement for social life. To achieve the trade-off between image privacy and image sharing, Sensitive Element Removing (SER) is studied in this paper.

Traditional processing methods like pixelation (Fan 2018), blurring, and masking (Li et al. 2017) significantly decrease the image quality after editing, which influences social usability. More importantly, with robust recognition methods (Oh et al. 2016) and reconstruction methods (Shen et al. 2019), image privacy cannot be guaranteed.

To make the modified image look natural with satisfactory social usability, image inpainting methods are adopted. For example, Uittenbogaard et al. (2019) proposed an inpainting framework to remove pedestrians and vehicles in street-view panoramas. However, without considering the original content, the results are sometimes illogical. In the task of SER, we take the original content into consideration to guarantee logical consistency. In addition, besides the target to naturally fill the specific regions, SER needs to make the content of the regions as different as possible after editing and keep the original content unrecognizable for visual perception.

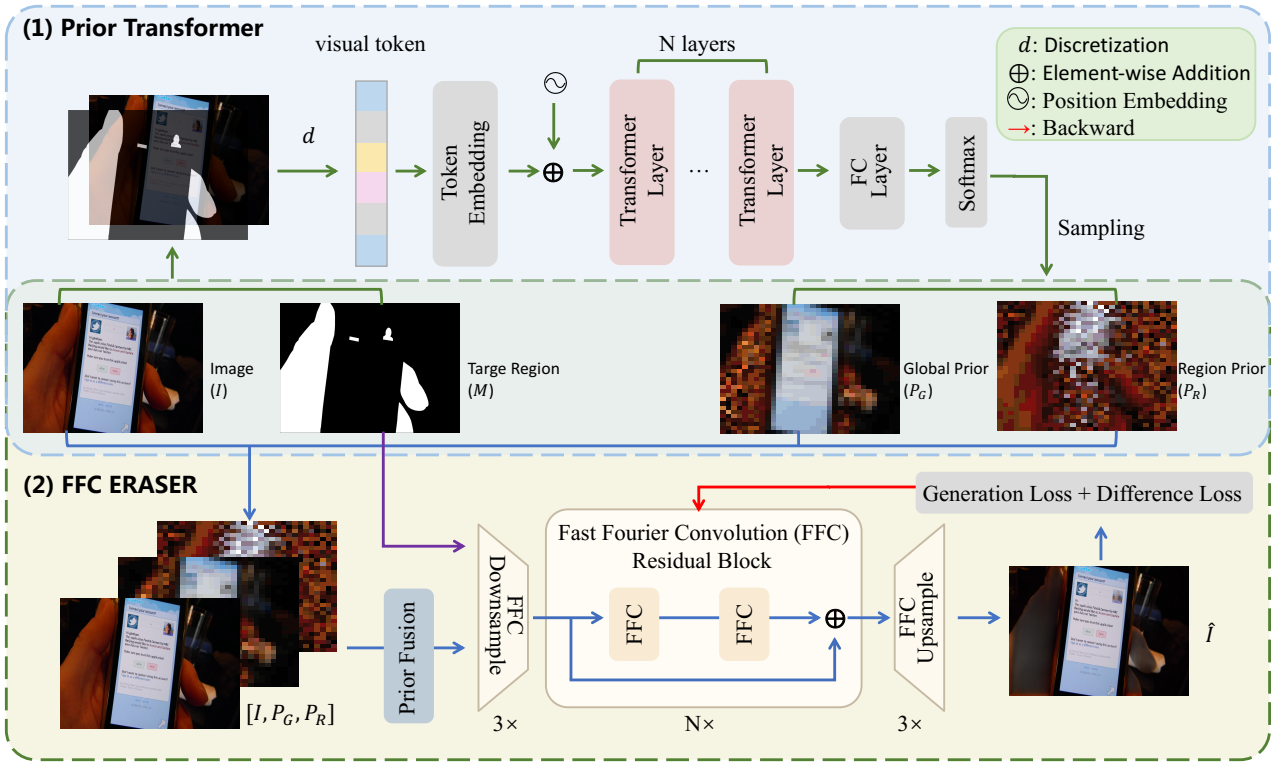

Figure 2: Overview of the AdvERsArial Sensitive Element Remover (ERASER), which contains two main parts. (1) Prior Transformer: With the input image I and the mask M that denotes the target region to be edited, ERASER considers the content of the target region M and builds the region prior P_R, as well as the content of the reserved region $\overline{M}$ and builds the global prior P_G. The P_R and P_G are sampled from the output of the Prior Transformer with Gibbs Sampling (Geman and Geman 1984), conditioned on $[I, M]$ and $[I, \overline{M}]$, respectively. (2) FFC ERASER: The image content and the image prior $[I, P_G, P_R]$ are fused and then combined with M. The fused information is downsampled and processed in both pixel and frequency domain with the Fast Fourier Convolution (FFC) blocks to edit the targeted regions. (3) While guaranteeing the image quality with Generation Loss, ERASER makes the edited region different with an additional Difference Loss.

Adversarial Sample for Image Privacy

Besides visual perception, another risk of image privacy is that some platforms may analyze the shared images automatically and recommend them to undesirable people. Adversarial methods are proposed to interfere with the adopted models and make them output wrong results. Such methods are adopted for image privacy preservation to make the images get rid of unauthorized automatic analysis and recognition of the platforms. For example, (Sun et al. 2018) proposed to add perturbation to the face, which makes the face detector invalid, Oh, Fritz, and Schiele (2017) focused on general objects and made object detectors invalid. In this paper, we only adopt the core idea. Instead of influencing the model output directly, we make the output feature from the backbone model different to interfere with the automatic recognition, while do not influence the quality and social usability of the image.

Frequency Domain Diagnose for Image Generation

Recently, frequency domain analyses have been studied in the field of image generation. For example, Dzanic, Shah, and Witherden (2020) discovered the discrepancies of high-frequency components in the Fourier spectrum between natural and generated images. Such findings demonstrate that current generative models are insufficient in the frequency domain, and some researchers adopt frequency-domain regularization to close the gap (Jiang et al. 2021). In this paper, instead of a total frequency-aware regularization as the supervision, we directly capture the feature characteristics in the frequency domain to improve the image quality with the Fast Fourier Convolution (Chi, Jiang, and Mu 2020).

ERASER

We propose ERASER to remove sensitive elements in the image for both image usability and image privacy, and the workflow is presented in Fig. 2. To take both global information and the target regions into consideration, we first adopt a transformer to build the region prior P_R conditioned on the image I and the mask M which indicates the regions of sensitive elements, and the global prior P_G conditioned on I and the mask $\overline{M}$ which indicates the rest regions. To improve the frequency-domain insufficiency of current methods, the

image, image prior, and the mask are processed with FFC blocks to edit the target sensitive elements. A Difference Loss is applied to make the content of the edited regions different from the original one.

Prior Transformer

We adopt the Transformer instead of convolution to construct the prior for two reasons. First, for the global prior P_G, we need to focus on the entire image, and Transformer has no limitation on the receptive field. Second, for the region prior P_R, we need to consider the correlation among the regions for a natural filling, and self-attention is very suitable for this task. Self-attention is the basic module in Transformer, and thus we adopt a Transformer directly.

Architecture

To build the image prior, we adopt the Transformer architecture proposed in (Wan et al. 2021) and is shown in Fig. 2 (1). First, to reduce the computational cost of multi-head attention (Vaswani et al. 2017), the image is compressed into a low-resolution version (i.e., 32×32) and then discretized with a visual vocabulary with a dimension of 512, and the image $I \in \mathcal{R}^{H \times W \times 3}$ is formatted as a discretized sequence $X = \{x_1, x_2, \cdots, x_{\mathbb{L}}\}$, where $\mathbb{L}$ is the length of $H \times W$.

With the Token Embedding Layer, X is projected into a vector $E_{img} \in \mathcal{R}^{\mathbb{L} \times d}$. Combined with the position embedding $E_{pos} \in \mathcal{R}^{\mathbb{L} \times d}$, $E = [E_{img}, E_{pos}] \in \mathcal{R}^{\mathbb{L} \times d}$ is used as the input for the Transformer layer. The Transformer only contains decoders following Radford et al. (2019), composed of several Multi-head Self Attention (MSA) layers. MSA calculates the attention result for each attention head $head_i$ and combines them as the result as follows:

$$\begin{aligned} head_i &= Attention(Q, K, V) \\ &= \text{softmax}\left(\frac{Q\mathbf{W}_Q^i \left(K\mathbf{W}_K^i\right)^T}{\sqrt{d}}\right)\left(V\mathbf{W}_V^i\right), \\ \text{MultiHead}(Q, K, V) &= \text{Concat}\left(\text{ head }_1, \ldots, \text{ head }_{\text{h}}\right) W^O, \end{aligned} \tag{1}$$

where d is the dimension of Q and K. $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are weights of fully-connected layers to obtain Q, K, V. h is the number of attention heads, Concat denotes the concatenation operation, and $\mathbf{W}_O$ is the weights of another fully-connected layer to fuse the feature from multi attention head.

For each Transformer layer, it processes the input as follows:

$$\begin{aligned} E_l' &= \text{LN}\left(\text{MSA}\left(E^{l-1}\right)\right) + E_{l-1}, \\ E_l &= \text{LN}\left(\text{MLP}\left(E_l'\right)\right) + E_l', \end{aligned} \tag{2}$$

where LN refers to Layer Normalization (Ba, Kiros, and Hinton 2016), MLP are several fully-connected layers. When used as self-attention, $Q = K = V = E$, and $\text{MSA}(E) = \text{MultiHead}(E, E, E)$.

To capture all the context, The Transformer adopted in this paper is not an auto-regressive one. After being processed with the fully-connected layer and softmax, the output of the last Transformer is projected to a per-element distribution over 512 elements (i.e., the dimension of visual vocabulary).

The Masked Language Model (MLM) is adopted for training following BERT (Devlin et al. 2019). Specifically, let X_Π denote the masked parts in X, and $X_{-\Pi}$ denotes the rest parts, the objective of MLM is to minimize the negative log-likelihood of X_Π conditioned on $X_{-\Pi}$:

$$L_{MLM} = \mathop{\mathbb{E}}_{X}\left[\frac{1}{K}\sum_{k=1}^{K} -\log p\left(x_{\pi_k} \mid X_{-\Pi}, \theta\right)\right], \tag{3}$$

where θ is the parameters of the Transformer.

Global and Region Prior Construction

Sampling from the output of the Prior Transformer can generate the image prior directly, but the results are not satisfactory due to the independence property. Therefore, Gibbs Sampling (Geman and Geman 1984) is adopted to iteratively sample tokens at different locations. Specifically, a patch is sampled from $p\left(x_{\pi_k} \mid X_{-\Pi}, X_{<\pi_k}, \theta\right)$, where $X_{<\pi_k}$ represents the sampled patches. By iteratively sampling the patches in a raster-scan manner, the image prior $X \in \mathbb{R}^{\mathbb{L} \times 3}$ is obtained.

To obtain the region prior P_R that contains information about the sensitive elements to be edited, the image I and the mask M that denotes the target regions are fed into the Prior Transformer. Similarly, to obtain the global prior P_G that contains information about the content to be reserved, the image I and the mask $\overline{M}$ that denotes the rest regions are processed with the Prior Transformer.

Two samples are visualized in Fig. 3. The input for the Prior Transformer is a low-resolution version (i.e.,32×32), and the prior has the same size as the input. The prior in 32×32 is resized in the figure for visualization. The low-resolution image prior contains main structural information and coarse textures, which is very suitable for SER to keep the main content and remove the details. Specifically, P_G describes the global information of the reserved content with a reasonable inference of the masked regions. The P_R describes the main structure and coarse textures of the regions to be edited like the people, the fingers, and part of the screen, with an inference of the surrounding context.

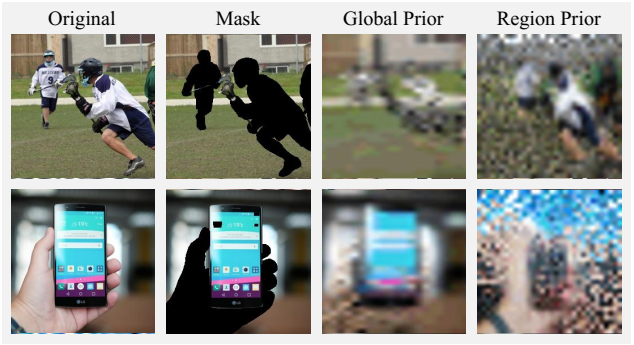

Figure 3: Visualization of the image prior. From left to right are original images, target regions denoted by masks, global prior P_G, and region prior P_R.

FFC ERASER

Although some researchers (Jiang et al. 2021; Fuoli, Van Gool, and Timofte 2021) adopt frequency-aware losses to regularize the generation methods, they only consider the final result in the frequency domain and the hidden states are ignored. To capture the feature characteristics beyond pixels, we adopt the Fast Fourier Convolution (FFC) to build FFC ERASER, and the workflow is presented in Fig. 2 (2).

By adopting FFC, FFC ERASER first edits the regions naturally, then make the content of the regions different. First, FFC ERASER fuses the image and the image prior $[I, P_G, P_R]$ to obtain the necessary information of regions to be edited and regions to be kept. The fused prior and the mask M are combined and downsampled with 3 FFC Downsample layers, and processed with several FFC Residual Blocks. At last, the processed information is processed with 3 FFC Upsample layers to reconstruct the image.

After training to converge with image quality losses, a Difference Loss is combined to make the content of the regions different in an adversarial way. The difference Loss is based on Perceptual Loss (Johnson, Alahi, and Fei-Fei 2016), which is the sum of the Content Loss and Style Loss (Gatys, Ecker, and Bethge 2016). ERASER minimizes the image quality losses while maximizing the Difference Loss to make difference. Considering that the losses for image quality contain a similar perceptual component implemented with ResNet (He et al. 2016), to make the Difference Loss not a subset of image quality loss, the Difference Loss is implemented with VGG (Simonyan and Zisserman 2015).

Experiments

To deal with the challenge that there is no direct dataset for SER, we propose a well-designed pipeline to train and evaluate ERASER. In this section, we introduce the pipeline and present the experimental results to answer the following evaluation questions:

EQ1 After integrating the Difference Loss, can ERASER still guarantee fine image quality for image editing?

EQ2 For SER, can ERASER achieve better image quality than other visual obfuscation methods?

EQ3 Does ERASER make the edited regions different enough after editing?

EQ4 Does ERASER indeed achieves the goal from the aspect of visual perception?

Experimental Setups

Datasets

Places2 (Zhou et al. 2017) is a scene recognition dataset that contains over 10 million images in 434 scene categories. The image-level annotations represent the entry-level of an environment like streets. Due to the abundant and diverse images, the dataset is widely used for image inpainting.

Visual Reduction (Orekondy, Fritz, and Schiele 2018) is an image privacy dataset with pixel label. The authors collected images with 24 types of sensitive elements and annotated them with segmentation masks. Overall, the pixel-labeled privacy dataset contains 8,473 images annotated with 47.6kk instances using 24 sensitive elements.

Implementation Details

There is no dataset of image pairs with and without sensitive elements, and thus ERASER cannot be trained directly. To tackle this challenge, the components of ERASER (i.e., Prior Transformer and FFC ERASER) were first trained with the large dataset for image inpainting (i.e., Places2) to generate high-quality images. Combined with the Difference Loss, ERASER was fine-tuned on the image privacy dataset with the segmentation of sensitive elements (i.e., Visual Reduction) to make the edited regions as different as possible.

Specifically, we first borrowed a well-trained visual Transformer from (Wan et al. 2021) to build the image prior P_G and P_R, with a size of 32×32. Then the Prior Fusion module fused $[I, P_G, P_R]$ with a 1×1 convolution layer, and output the fused prior with the dimension of $256 \times 256 \times 3$. There were 3 FFC Downsample layers, 18 FFC Residual Blocks, and 3 FFC Upsample layers in FFC ERASER. The input image and the mask were resized to 256×256 during training, with the data augmentation of horizontal flip. The output image has the same size as the input one (i.e., 256×256). During inference, there is no limitation for the input size of ERASER.

The Blur and Pixelation were implemented by Pillow with a radius of 8. The pretrained parameters of the Transformer and the FFC blocks were borrowed from (Wan et al. 2021) and (Suvorov et al. 2022), respectively.

Compared Methods

We conducted several obfuscation methods on Visual Reduction. We first conducted three traditional image processing methods, blurring, pixelating, and masking, to remove the sensitive elements. Although such methods cannot guarantee image privacy, we provide their results on image quality and content difference for reference. We further compare with two state-of-the-art image inpainting methods (i.e., ICT (Wan et al. 2021) and LaMa (Suvorov et al. 2022)), which have a similar target to SER.

Quantitative Analyses

Comparison with Inpainting Methods (EQ1)

Although ERASER is designed for SER rather than image inpainting, the input data (i.e., image and mask) for inpainting is also valid for ERASER. Therefore, we ran ERASER on the image inpainting dataset first to validate that ERASER can generate images with good quality, and the results are presented in Table 1.

All the results are based on the Places2 test set. The results of ERASER are from the final model fine-tuned with the Difference Loss, and the results of image inpainting methods are reported by (Wan et al. 2021) and (Suvorov et al. 2022). We observe that the performance of ERASER is comparable with other models in the task of image inpainting and just a bit lower than the state-of-the-art ones. The performance on the large dataset demonstrates that ERASER can generate images with good quality. The Difference Loss does not influence the image quality too much.

Effectiveness of Image Quality for SER (EQ2)

To validate the image quality of SER methods, experiments were conducted on the previously introduced methods, and

Method	PSNR↑	SSIM↑	MAE↓
DFv2 (Yu et al. 2019)	25.692	0.834	0.0280
EC (Nazeri et al. 2019)	25.510	0.831	0.0293
PIC (Zheng, Cham, and Cai 2019)	25.035	0.806	0.0315
MED (Liu et al. 2020b)	25.632	0.827	0.0291
ICT (Wan et al. 2021)	**25.982**	0.839	0.0254
LAMA (Suvorov et al. 2022)	24.947	**0.857**	**0.0245**
ERASER	24.085	0.845	0.0285

Table 1: Comparison with inpainting methods on Places2. "↑" indicates higher is better, while "↓" indicates lower is better. The best results in each column are boldfaced.

the results are presented in Table 2. The images edited by Blur and Pixelation can even be recovered, and thus the image quality is higher than most other methods. Compared with Mask and the other two image inpainting methods, ERASER achieved the highest image quality. Compared with Blur and Pixelation, the PSNR of ERASER is comparable, while the SSIM of ERASER is even higher. Generally speaking, ERASER achieved the best image quality among the valid SER methods.

Method	SSIM↑	PSNR↑	FID↓	LPIPS↓
Mask	0.622	14.232	75.57	0.378
Blur	0.841	27.777	26.02	0.098
Pixelation	0.847	29.199	21.40	0.084
ICT	0.646	18.261	99.32	0.190
LAMA	0.694	18.866	**63.23**	**0.119**
ERASER	**0.862**	**25.592**	65.35	0.125

Table 2: Image quality comparison with other sensitive element removing methods. "↑" indicates higher is better, while "↓" indicates lower is better.

Effectiveness of Content Difference (EQ3)

To validate the differences of the regions after editing, we adopted several metrics to evaluate the difference, including metrics in pixel-level (MAE), feature-level (Feature Matching (FM), Perceptual Loss from VGG (PL-V) and ResNet (PL-R)), and classification-level (Adversarial Loss (Adv)), and the results are presented in Table 3. In general, ERASER achieved a robust performance in all the metrics, while most other methods are not robust under certain metrics for the difference. Such results prove that ERASER can interfere with the recognition of models like VGG (Simonyan and Zisserman 2015) and ResNet (He et al. 2016).

Specifically, for all the metrics, the differences of Blur and Pixelation are not very significant, which proves that these methods cannot guarantee image privacy. For Mask, the difference for Adv, PL-V, and PL-R is not significant, which is consistent with the finding in (Oh et al. 2016) that the identity is still recognizable with content masked. The LAMA (Suvorov et al. 2022) has similar performance compared with ERASER but was achieved with the expense of image quality degrading as shown in Table 2.

Method	MAE↑	FM↑	Adv↑	PL-V↑	PL-R↑
Mask	0.179	**0.301**	0.633	0.040	0.730
Blur	0.023	0.083	0.632	0.041	0.052
Pixelation	0.018	0.074	0.631	0.040	0.055
ICT	**0.237**	**0.266**	13.035	49.162	**65.056**
LAMA	0.178	0.206	**29.863**	131.789	55.774
ERASER	0.198	0.215	29.860	**131.819**	55.937

Table 3: Content difference comparison with other sensitive element removing methods. MAE is obtained from pixel values of the image, while others are obtained from deep features of neural networks. A higher value means a larger difference for all metrics. FM, Adv, PL-V, and PL-R indicate Feature Matching, Adversarial Loss, Perceptual Loss obtained from VGG and ResNet, respectively.

Summary Overall, the Difference Loss achieves the goal of making the content in target regions different (EQ3), and the image quality is better than other SER methods (EQ2). At the same time, it does not influence the image quality too much, and the performance of image inpainting is still comparable with state-of-the-art methods (EQ1).

Qualitative Analyses (EQ4)

The above analyses prove that ERASER achieves the goal in the aspect of quantitative metrics. To demonstrate that ERASER indeed removes the sensitive elements while guaranteeing the image quality from the aspect of visual perception, we conduct qualitative analyses. We have presented several samples in Fig. 1, and additional samples are presented in Fig. 4. To remove the sensitive elements in the original images (first row) denoted by masks (second row), we adopt both state-of-the-art image inpainting method (Wan et al. 2021) (third row) and ERASER (bottom row).

We first analyze the performance with the sensitive elements of objects, which cover various types of sensitive information. The results are presented on the left of Fig. 4. The sensitive elements include username, avatar, fingerprint (first column); license plate (second column); and landmarks (the third column). In general, both (Wan et al. 2021) and ERASER removes the sensitive elements from the visual perceptual aspect, but the results of (Wan et al. 2021) are not very natural and decrease social usability.

People are another kind of sensitive element that often appear in the images. The corresponding results are presented on the right of Fig. 4. The images are related to private scenarios (the fourth column), and sensitive scenarios like protests (the last column). Similarly, (Wan et al. 2021) removes the people but causes obvious artifacts. ERASER generates a harmonious shadow to replace the people.

Discussion

In general, ERASER achieves the goal to remove sensitive elements from the visual perceptive aspect. More importantly, ERASER achieves the trade-off between image privacy and image usability. The images are relatively natural even with large edited regions.

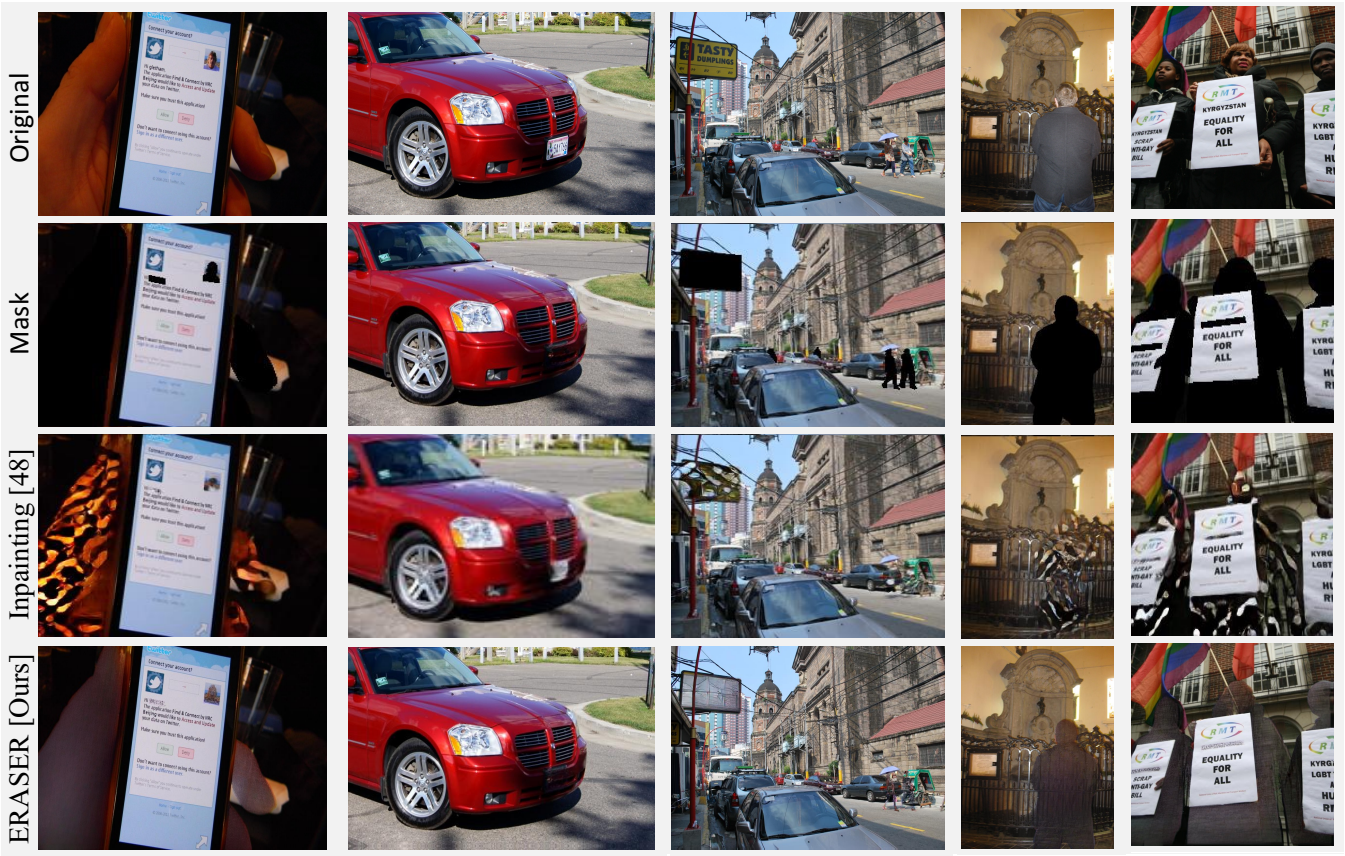

Figure 4: Results of images with sensitive elements. From top to bottom are original images, sensitive elements denoted by black masks, results of image inpainting (Wan et al. 2021), and results of ERASER. (Zoom in for the best of view.)

Figure 5: Cases that ERASER cannot well handled, including people with complex postures and credentials without detailed annotation of sensitive regions.

On the other hand, ERASER has some limitations. First, whether the person is replaced with a shadow or with the background is not controllable. Second, for people with complex postures (first row in Fig. 5), ERASER cannot fill the regions very well. As there are many human-centered generative methods, we suggest using more targeted models instead, such as face reenactment (Song et al. 2021) and appearance transfer (Zanfir et al. 2018). Third, for credentials, certificates, and cards (second row in Fig. 5), the whole of them are annotated as sensitive in the Visual Reduction dataset and are fully edited without usability. The fine-grained annotation will help deal with such scenarios.

In this paper, the sensitive regions need to be specified by users. Some researchers try to locate sensitive elements in the images automatically (Yu et al. 2018; Yang et al. 2022). Therefore, a promising future work is to combine such methods with sensitive element removing method and build end-to-end image privacy preserving frameworks.

Conclusion

In this paper, we propose the task of Sensitive Element Removing that naturally fills specific regions, and makes the modified regions different for both human perception and model recognition to preserve image privacy. We propose ERASER and corresponding training and evaluating pipelines without enough image pairs of ground truth. ERASER is designed to capture global and region structures in both pixel and frequency domains to make the modified image more natural, then combined a Difference Loss to make the regions different. Both quantitative and qualitative analyses demonstrate that ERASER achieves the balance between image privacy and image usability.

Acknowledgements

The corresponding author is Juan Cao. This work was supported by the National Key Research and Development Program of China (2021AAA0140203), the National Natural Science Foundation of China (62203425), and the Project of Chinese Academy of Sciences (E141020).

References

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Chi, L.; Jiang, B.; and Mu, Y. 2020. Fast fourier convolution. In *Advances in Neural Information Processing Systems*, volume 33, 4479–4488.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186.

Dzanic, T.; Shah, K.; and Witherden, F. D. 2020. Fourier Spectrum Discrepancies in Deep Network Generated Images. In *Advances in Neural Information Processing Systems*.

Equifax. 2020. Protect against identity theft when sharing photos online. https://www.equifax.co.uk/resources/identity-protection/protect-against-identity-theft-when-sharing-photos-online.html/. Accessed: November, 2020.

Fan, L. 2018. Image pixelization with differential privacy. In *IFIP Annual Conference on Data and Applications Security and Privacy*, 148–162.

Fuoli, D.; Van Gool, L.; and Timofte, R. 2021. Fourier Space Losses for Efficient Perceptual Image Super-Resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2360–2369.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2414–2423.

Geman, S.; and Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6): 721–741.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hu, M.; Yang, Y.; Shen, F.; Xie, N.; Hong, R.; and Shen, H. T. 2018. Collective reconstructive embeddings for cross-modal hashing. *IEEE Transactions on Image Processing*, 28(6): 2770–2784.

Jiang, L.; Dai, B.; Wu, W.; and Loy, C. C. 2021. Focal Frequency Loss for Image Reconstruction and Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 694–711.

Li, Y.; Vishwamitra, N.; Knijnenburg, B. P.; Hu, H.; and Caine, K. 2017. Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1343–1351.

Liu, C.; Zhu, T.; Zhang, J.; and Zhou, W. 2020a. Privacy Intelligence: A Survey on Image Sharing on Online Social Networks. *arXiv preprint arXiv:2008.12199*.

Liu, H.; Jiang, B.; Song, Y.; Huang, W.; and Yang, C. 2020b. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *European Conference on Computer Vision*, 725–741.

Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.; and Ebrahimi, M. 2019. EdgeConnect: Structure Guided Image Inpainting using Edge Prediction. In *The IEEE International Conference on Computer Vision Workshops*.

Oh, S. J.; Benenson, R.; Fritz, M.; and Schiele, B. 2016. Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision*, 19–35.

Oh, S. J.; Fritz, M.; and Schiele, B. 2017. Adversarial image perturbation for privacy protection a game theory perspective. In *Proceedings of the IEEE International Conference on Computer Vision*, 1491–1500.

Orekondy, T.; Fritz, M.; and Schiele, B. 2018. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8466–8475.

Orekondy, T.; Schiele, B.; and Fritz, M. 2017. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE International Conference on Computer Vision*, 3686–3695.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Shen, L.; Hong, R.; Zhang, H.; Zhang, H.; and Wang, M. 2019. Single-shot Semantic Image Inpainting with Densely Connected Generative Networks. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1861–1869.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

Solsman, J. E. 2020. Deepfake bot on Telegram is violating women by forging nudes from regular pics. https://www.cnet.com/news/deepfake-bot-on-telegram-is-violating-women-by-forging-nudes-from-regular-pics/. Accessed: October, 2020.

Song, L.; Wu, W.; Fu, C.; Qian, C.; Loy, C. C.; and He, R. 2021. Everything's Talkin': Pareidolia Face Reenactment. In *Proceedings of the IEEE International Conference on Computer Vision*.

Song, Y.; Yang, C.; Lin, Z.; Liu, X.; Huang, Q.; Li, H.; and Kuo, C.-C. J. 2018. Contextual-based image inpainting: Infer, match, and translate. In *Proceedings of the European Conference on Computer Vision*, 3–19.

Sun, Q.; Ma, L.; Oh, S. J.; Van Gool, L.; Schiele, B.; and Fritz, M. 2018. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5050–5059.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust Large Mask Inpainting with Fourier Convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2149–2159.

Tonge, A.; and Caragea, C. 2019. Dynamic deep multimodal fusion for image privacy prediction. In *The World Wide Web Conference*, 1829–1840.

Uittenbogaard, R.; Sebastian, C.; Vijverberg, J.; Boom, B.; Gavrila, D. M.; et al. 2019. Privacy protection in street-view panoramas using depth and multi-view imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10581–10590.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Wan, Z.; Zhang, J.; Chen, D.; and Liao, J. 2021. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4692–4701.

Xiong, W.; Yu, J.; Lin, Z.; Yang, J.; Lu, X.; Barnes, C.; and Luo, J. 2019. Foreground-aware image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5840–5848.

Yang, G.; Cao, J.; Chen, Z.; Guo, J.; and Li, J. 2020. Graph-Based Neural Networks for Explainable Image Privacy Inference. *Pattern Recognition*, 107360.

Yang, G.; Cao, J.; Sheng, Q.; Qi, P.; Li, X.; and Li, J. 2022. DRAG: Dynamic Region-Aware GCN for Privacy-Leaking Image Detection. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*.

Yang, Y.; Zhou, J.; Ai, J.; Bin, Y.; Hanjalic, A.; Shen, H. T.; and Ji, Y. 2018. Video captioning by adversarial LSTM. *IEEE Transactions on Image Processing*, 27(11): 5600–5611.

Yu, J.; Kuang, Z.; Zhang, B.; Zhang, W.; Lin, D.; and Fan, J. 2018. Leveraging content sensitiveness and user trustworthiness to recommend fine-grained privacy settings for social image sharing. *IEEE Transactions on Information Forensics and Security*, 13(5): 1317–1332.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2019. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4471–4480.

Zanfir, M.; Popa, A.; Zanfir, A.; and Sminchisescu, C. 2018. Human Appearance Transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5391–5399.

Zerr, S.; Siersdorfer, S.; Hare, J.; and Demidova, E. 2012. Privacy-aware image classification and search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 35–44.

Zheng, C.; Cham, T.-J.; and Cai, J. 2019. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1438–1447.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6): 1452–1464.