SCENE PRIVACY PROTECTION

Chau Yi Li, Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, Riccardo Mazzon, Andrea Cavallaro

Centre for Intelligent Sensing, Queen Mary University of London, UK

ABSTRACT

Images shared on social media are routinely analysed by classifiers for content annotation and user profiling. These automatic inferences reveal to the service provider sensitive information that a naive user might want to keep private. To address this problem, we present a method designed to distort the image data so as to hinder the inference of a classifier without affecting the utility for social media users. The proposed approach is based on the Fast Gradient Sign Method (FGSM) and limits the likelihood that automatic inference can expose the true class of a distorted image. Experimental results on a scene classification task show that the proposed method, *private* FGSM, achieves a desirable trade-off between the drop in classification accuracy and the distortion on the private classes of the Places365-Standard dataset using ResNet50. The classifier is misled 94.40% of the times in the top-5 classes with only a small average reduction of three image quality measures (SSIM, PSNR, BRISQUE).

***Index Terms*—** Privacy; Adversarial Images; Fast Gradient Sign Method; Image Quality.

1. INTRODUCTION

Routine large-scale inference on images shared on social media reveals information (image content) that contributes to create detailed user profiles, which can then be used for targeted commercial or political advertising. As images may capture details that are sensitive (private) for a user, a privacy violation may occur when a classifier infers, without user consent, sensitive information from an image. We therefore aim to protect the private content of images that a user shares with other users from undesirable *automatic* inference.

Classifiers can infer from images the presence of people, their age, gender, clothing style, their relationship as well as the scene class depicted in the image [1]. Traditional methods for visual privacy protection distort the appearance of sensitive image regions (e.g. faces) to make them unrecognisable using redactions [2], cartooning [3], pixelation [4], single or multiple blurs [5, 6], false colours [7], scrambling [8] or warping [9]. Moreover, using deep learning pipelines, face regions can be de-identified while preserving their original facial expression [10].

The key properties of an ideal method for privacy protection against the automatic inference of sensitive information are (i) to maintain the fidelity (utility) of images so that people cannot notice the distortion; (ii) to conceal the distortion so that an algorithm cannot detect it; and (iii) to prevent the deduction of a mapping between the true class and the class assigned by the classifier to the distorted image. In summary, the impact of the distortion on a protected image should be *unnoticeable*, *undetectable* and *irreversible*.

The first three authors contributed equally. Andrea Cavallaro wishes to thank the Alan Turing Institute (EP/N510129/1), which is funded by the EPSRC, for its support through the project PRIMULA.

To this end, we exploit the knowledge that relatively small perturbations in an image [11] can mislead specific object detectors [12] and image classifiers [11, 13, 14]. For example, if the true class is known, the Fast Gradient Sign Method (FGSM) [15] generates distorted images that are most likely to be classified as the nearest incorrect class [16]. To improve success in misleading the classifier [17], FGSM can be re-applied iteratively until a desired misclassification probability is reached [18]. Variations of FGSM produce distorted images that are (mis)classified as the least likely class [18] or as a randomly selected class (except the most likely one) [19]. However, these adversarial methods have the following limitations for privacy protection: FGSM and iterative FGSM require the availability of the true class of an image [16], thus limiting their applicability as a user would need to declare the true class for each image to be uploaded; least-likely FGSM is reversible as the true class of the image can be recovered with high probability; and random FGSM may select the true class as the original image is classified correctly only about half of the times, as we will see in this paper.

We address the above limitations with *private* FGSM, a privacy-protection method that achieves diversity of the selected target class and reduces the likelihood that the mapping between the original and the target class can be deduced. We achieve diversity by picking the target class from an adaptive subset of classes that most likely does *not* include the class to be protected. This subset is defined based on the inference probabilities of the classifier on the image to be protected. We validate *private* FGSM on the task of preserving the privacy of a scene, such as a place of cult or a hospital, and evaluate the irreversibility and quality of the adversarial image generated for the ResNet50 classifier on the Places365-Standard dataset [20].

2. PRIVATE FAST GRADIENT SIGN METHOD

Let $\mathbf{x}$ be an image and $\hat{y}_i$ be the true class label of the scene type depicted in $\mathbf{x}$. The label $\hat{y}_i$ belongs to a set of D scene classes, $\{\hat{y}_1, ..., \hat{y}_i, ..., \hat{y}_D\}$. Applying a multi-class classifier M to $\mathbf{x}$ generates the D-dimensional one-hot vector $\mathbf{y}$:

$$\mathbf{y} = M(\mathbf{x}), \tag{1}$$

where $\mathbf{y} = (y_1, ..., y_i, ..., y_D)$ results from a decision on the D-dimensional vector $\mathbf{p} = (p_1, ..., p_i, ..., p_D)$, whose element p_i is the probability that $\mathbf{x}$ depicts the scene class y_i:

$$p_i = p(y_i|\mathbf{x}). \tag{2}$$

We aim to define a transformation T such that $\dot{\mathbf{x}} = T(\mathbf{x})$ induces M to classify the image with a different scene label:

$$\mathbf{y} \neq M(\dot{\mathbf{x}}). \tag{3}$$

The distortion applied to the image $\mathbf{x}$ by T should be minimal ($\|\dot{\mathbf{x}} - \mathbf{x}\| \rightarrow 0$) so that the transformation is *unnoticeable*. Moreover, T should be *irreversible* so that the true class, $\hat{y}_i$, cannot be

deduced from the predicted class $M(\dot{\mathbf{x}})$ or from the distribution of the probabilities of the predicted classes. Let us define T as follows:

$$\dot{\mathbf{x}} = T(\mathbf{x}) = \mathbf{x} + \delta^*_{\mathbf{x}}, \tag{4}$$

where $\delta^*_{\mathbf{x}}$ is adversarial noise that can be generated as:

$$\delta^*_{\mathbf{x}} = \arg\max_{\delta_{\mathbf{x}}} J_M(\theta, \mathbf{x} + \delta_{\mathbf{x}}, \mathbf{y}), \tag{5}$$

where J_M is the cost function used in training to estimate the parameters θ of classifier M. Eq. 5 maximises the error for the originally predicted class and has no closed-form solution if J_M is non-convex [18]. When J_M is the cross-entropy function and the parameters of the classifier are known, FGSM can generate adversarial noise to induce the classifier to select a specific class label (e.g. targeted least-likely FGSM [18], targeted random FGSM [19]) or to increase the misclassification probability (e.g. non-targeted FGSM [15], non-targeted iterative FGSM [18]).

FGSM [15] solves Eq. 5 by linearising J_M around θ using the true class $\hat{y}_i$ represented by the one-hot vector $\hat{\mathbf{y}}$:

$$\dot{\mathbf{x}} = \mathbf{x} + \epsilon \, \mathrm{sign}(\nabla_{\mathbf{x}} J_M(\theta, \mathbf{x}, \hat{\mathbf{y}})), \tag{6}$$

where ϵ controls the magnitude of the perturbation and $\nabla_{\mathbf{x}} J_M$ is the gradient of the cost function J_M with respect to $\mathbf{x}$. FGSM requires only one inference and one backpropagation of M and therefore is fast, but does not guarantee the misclassification of the image [17].

Iterative FGSM [18] extends FGSM by generating adversarial noise iteratively until a desired (mis)classification probability or a maximum number of iterations is reached. The final $\dot{\mathbf{x}} = \dot{\mathbf{x}}_N$ is obtained as

$$\dot{\mathbf{x}}_N = \dot{\mathbf{x}}_{N-1} + \epsilon \, \mathrm{sign}(\nabla_{\mathbf{x}} J_M(\theta, \dot{\mathbf{x}}_{N-1}, \hat{\mathbf{y}})), \tag{7}$$

from the initialisation $\dot{\mathbf{x}}_0 = \mathbf{x}$. For example, our experiments reached convergence with fewer than 20 iterations, on average.

FGSM and iterative FGSM need to know the true class [18], which may be unavailable or impractical to generate in real-world applications. To address this limitation, least-likely FGSM [18] forces the transformation T to target the least-likely class. If $\overline{\mathbf{y}}$ is the least-likely D-dimensional one-hot vector, whose elements are

$$\overline{y}_i = \begin{cases} 1 & i = \arg\min_{j=1\ldots D} p(y_j|\mathbf{x}) \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

least-likely FGSM generates $\dot{\mathbf{x}} = \dot{\mathbf{x}}_N$ iteratively, from the initialisation $\dot{\mathbf{x}}_0 = \mathbf{x}$, as

$$\dot{\mathbf{x}}_N = \dot{\mathbf{x}}_{N-1} - \epsilon \, \mathrm{sign}(\nabla_{\mathbf{x}} J_M(\theta, \dot{\mathbf{x}}_{N-1}, \overline{\mathbf{y}})), \tag{9}$$

by increasing the probability of predicting $\overline{\mathbf{y}}$ until a desired classification probability or a maximum number of iterations is reached. However, selecting always as target class the least-likely one can be exploited to deduce the true class, thus compromising *irreversibility*.

To overcome this problem, random FGSM [19] modifies least-likely FGSM by selecting randomly the target class, $\tilde{y}$, from the set of all possible classes except the most-likely class. However, as the top-1 accuracy of classifiers may be low (see Table 1) random FGSM may select the true class as the target class.

To achieve *irreversibility* with a high misclassification rate, we propose *private* FGSM, a targeted and iterative FGSM, which generates adversarial images by adaptively targeting a class, $\tilde{y}$, selected as a function of the classification probability vector $\mathbf{p}$. We obtain a high misclassification rate by leveraging the fact that the true class is often among the classes with the highest cumulative probabilities.

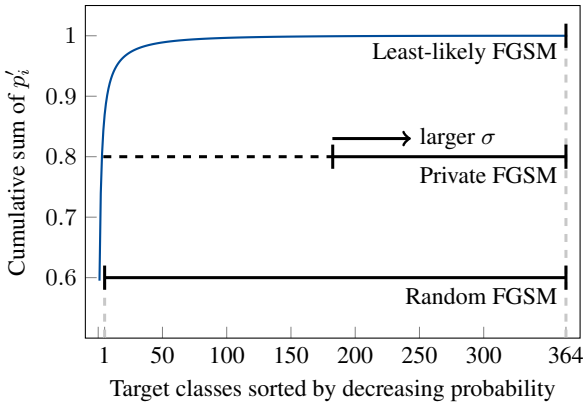

Fig. 1. Comparison of the target class selection strategies for least-likely FGSM, random FGSM and *private* FGSM. Least-likely FGSM selects always the least-likely class. Random FGSM chooses from the set of all classes but the most-likely one (e.g. 364 classes in the Places365-Standard dataset). *Private* FGSM chooses from a set whose number of classes is determined by a minimum cumulative probability defined by σ (see Eq. 10).

Let $\mathbf{p}' = (p'_1, ..., p'_D)$ contain the elements of $\mathbf{p}$ sorted in descending order. *Private* FGSM selects $\tilde{y}$ randomly from the subset of classes whose cumulative probability exceeds a threshold $\sigma \in [0, 1]$:

$$\tilde{y} = R\left(\left\{y_j : \sum_{i=1}^{j-1} p'_i > \sigma\right\}\right), \tag{10}$$

where R is a function that picks randomly one class label from the input set and σ controls the number of classes to pick $\tilde{y}$ from: the larger σ, the smaller the subset of target classes (see Fig. 1). The protected image $\dot{\mathbf{x}} = \dot{\mathbf{x}}_N$ is generated iteratively, starting from $\dot{\mathbf{x}}_0 = \mathbf{x}$, as

$$\dot{\mathbf{x}}_N = \dot{\mathbf{x}}_{N-1} - \epsilon \, \mathrm{sign}(\nabla_{\mathbf{x}} J_M(\theta, \dot{\mathbf{x}}_{N-1}, \tilde{\mathbf{y}})), \tag{11}$$

by increasing the probability of predicting $\tilde{\mathbf{y}}$ until a desired classification probability or a maximum number of iterations is reached.

Unlike least-likely FGSM [18], *private* FGSM increases diversity to favour *irreversibility* as the target class is randomly selected among the subset of classes that most likely does *not* contain the class to be protected. As example of mapping from the true class to the target class in the Places365-Standard dataset [20], least-likely FGSM maps the class *aqueduct* to *operating room* 38% of the times and iterative FGSM maps the class *bus interior* to *train interior* 38% of the times, thus making these methods less suitable for privacy protection. Instead, the highest frequency of a class consistently mapped to a target class is 8% with random FGSM and 6% with *private* FGSM.

Fig. 2 shows two examples of transformed images with *private* FGSM and their selected classes $\tilde{y}$.

3. VALIDATION

3.1. Experimental setup

We compare the proposed method, *private FGSM* (P-FGSM), with FGSM [15], iterative FGSM (N-FGSM) [18], least-likely FGSM (L-

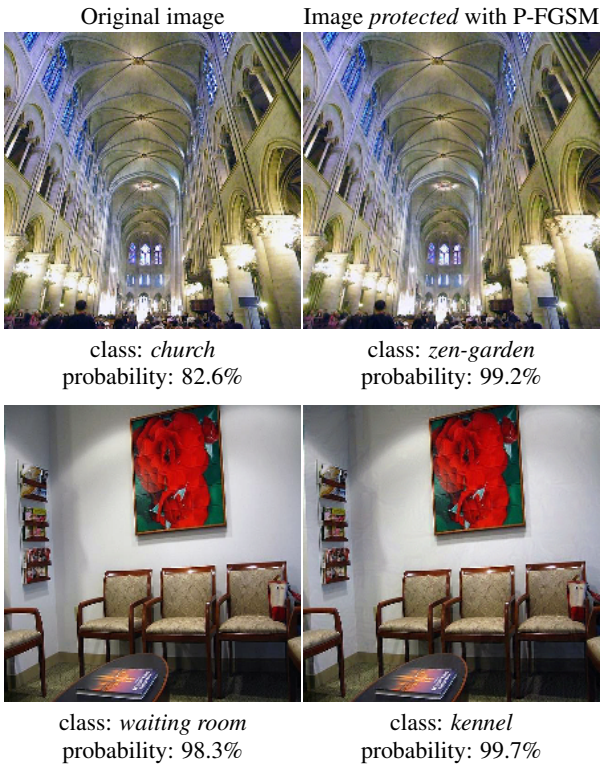

Fig. 2. Examples of two images (left column) transformed into their protected versions (right column) with the proposed *private* FGSM (P-FGSM).

FGSM) [18], random FGSM (R-FGSM) [19] and with a set of baseline adversarial methods: a generic sign of Gaussian (SoG), with zero mean and unit standard deviation, multiplied by ϵ, added for 1 or 8 iterations; the FGSM noise (see Eq. 6), added for 6 iterations; and the R-FGSM noise, added for 1 or 3 iterations. The number of iterations is selected to obtain a classification accuracy comparable to that of the other methods, when possible. Images are clipped at each iteration. For SoG, R-FGSM, L-FGSM and P-FGSM, we select $\epsilon = 0.007$, which is associated with the smallest pixel variation in an 8-bit image; the desired classification probability threshold is 0.99; and the maximum number of iterations is 50 to limit the amount of adversarial noise introduced. For P-FGSM, we select $\sigma = 0.99$ as trade-off, on the training dataset, between privacy protection (misclassification) and irreversibility.

3.2. Dataset and classifier

We use as *scene privacy* dataset a subset of the validation set of the Places365-Standard dataset [20] defined[1] in the Mediaeval 2018 Pixel Privacy Challenge [21]. The Places365-Standard dataset has over 1.8 million images of 365 scene classes, divided into training, validation and testing sets. The training and testing datasets of the Challenge contain each 3,000 images that are part of 60 private classes, with 50 images per class. While the testing images include the 60 private scene classes only, the target class can be selected as any of the scene classes of the Places365 dataset.

We use ResNet50 as *multi-class classifier* that was selected by the Mediaeval 2018 Pixel Privacy Challenge [21]. We apply a bilinear interpolation to downsize the original images to 224×224 pixels instead of using the re-sized images of the Places365-Standard dataset (256×256 pixels, cropped to 224×224 pixels). Our solution is preferable for both classification performance and image quality.

3.3. Evaluation measures

To evaluate the extent to which a transformation T can protect the privacy of the content of an image while maintaining its utility, we consider the classification (in)accuracy, the *irreversibility*, and the visual quality of the transformed image.

The classification accuracy quantifies the ability of a transformed (protected) image to mislead the classifier. We evaluate at which rank the transformed image is correctly classified by M. The lower the accuracy in top ranks, the better the protection.

An irreversible transformation should have no bias towards a particular target class for a given true class, hence the distribution of the target class should ideally be indistinguishable from a uniform distribution. We quantify *irreversibility* with the Euclidean distance between the discrete uniform distribution and the average discrete distribution of the target class of all the classes under consideration. The smaller the distance, the higher the *irreversibility*.

To quantify the extent to which the protection is *unnoticeable* we use three quality measures, namely the Structural SIMilarity (SSIM) index [22], the Peak-Signal-to-Noise Ratio (PSNR), and the Blind Referenceless Image Spatial Quality Evaluator (BRISQUE) [23]. SSIM is a full-reference measure that quantifies the structure preservation in image windows: the higher the SSIM, the better the image quality. The Peak-Signal-to-Noise Ratio (PSNR) is another full-reference measure that quantifies the pixel-by-pixel difference between two images: the higher the PSNR, the better the image quality. Finally, BRISQUE is a no-reference measure that quantifies distortions and unnaturalness in an image: the lower BRISQUE, the better the image quality.

3.4. Discussion

Table 1 shows the top-1 and top-5 classification accuracy with ResNet50, and SSIM, PSNR and BRISQUE as mean and standard deviation for all the images. FGSM, N-FGSM and the baseline methods either do not mislead the classifier (e.g. high accuracy of N-FGSM) or considerably drop in visual quality (e.g. R-FGSM with 3 iterations). The images transformed by P-FGSM, N-FGSM, L-FGSM and R-FGSM have comparable visual quality.

R-FGSM obtains a low classification accuracy with 0.17% and 7.00% for top-1 and top-5 accuracy, respectively. P-FGSM and L-FGSM achieve the lowest top-1 accuracy by always misleading the classifier, whereas L-FGSM obtains the lowest top-5 accuracy. However, L-FGSM is less *irreversible* than N-FGSM, R-FGSM and P-FGSM.

[1]The subset of sensitive classes includes scenes that may require, for various reasons, privacy protection, such as *army-base, bathroom, bedchamber, bedroom, church (indoor and outdoor), hospital and hospital-room, nursing-home, pharmacy, sauna, shower, swimming pool (indoor and outdoor), jacuzzi (indoor), temple (Asia)*; as well as scenes that would disclose private personal information, such as *airplane-cabin, airport-terminal, amusement-park, aqueduct, bank-vault, bar, beach, beach-house, beer-garden, beer-hall, berth, bullring, bus-interior, bus-station/indoor, campsite, car-interior, castle, catacomb, chalet, childs-room, classroom, closet, coast, discotheque, dorm-room, drugstore, gymnasium (indoor), home-office, kindergarten-classroom, locker-room, mosque (outdoor), playground, playroom, pub (indoor), sandbox, schoolhouse, ski-resort, ski-slope, slum, swimming-hole, train-interior, train-station/platform, tree-house, and waiting-room*.

Table 1. Classification accuracy on the private images of the testing subset of Places365-Standard [21] and visual quality scores (with standard deviation). Classifier: ResNet50. KEY – T1: top-1 accuracy (%); T5: top-5 accuracy (%); Orig.: Original image bilinearly downsampled to 224×224; (XI): noise added for X iterations; SoG: Sign of Gaussian; FG: Fast Gradient Sign Method (FGSM); SSIM: Structural Similarity Index; PSNR: Peak Signal to Noise Ratio; BRISQUE [23]; ↓: the lower, the better; ↑: the higher, the better.

Method	T1 ↓	T5 ↓	SSIM ↑	PSNR ↑	BRISQUE ↓
Orig.	56.40	86.47	-	-	26.71 (8.66)
SoG (1I)	56.43	86.53	.99 (.01)	42.16 (0.11)	24.90 (8.84)
SoG (8I)	1.36	5.03	.26 (.10)	12.85 (0.43)	46.63 (3.32)
FG	10.56	46.12	.99 (.01)	42.15 (0.13)	25.11 (8.42)
FG (6I)	7.93	17.20	.83 (.09)	27.65 (0.30)	41.32 (4.48)
R-FG (1I)	16.93	47.13	.99 (.01)	42.15 (0.12)	25.08 (8.43)
R-FG (3I)	0.37	5.87	.94 (.03)	34.23 (0.17)	33.40 (7.65)
N-FG	8.83	23.00	.98 (.02)	40.62 (4.75)	24.16 (8.31)
R-FG	0.17	7.00	.99 (.01)	40.24 (2.87)	23.99 (8.29)
L-FG (*)	0.00	0.17	.99 (.01)	38.08 (2.30)	23.67 (8.36)
P-FG	0.00	5.60	.99 (.01)	39.99 (2.72)	23.85 (8.28)

(*) considerably less *irreversible* than N-FG, R-FG and P-FG

Table 2. Euclidean distance between the uniform distribution and the distribution of the target class selection with four methods. The lower the distance, the higher the irreversibility.

Method	Distance
N-FGSM	0.2298
R-FGSM	0.1414
L-FGSM	0.2805
P-FGSM	0.1416

Table 2 shows the Euclidean distance between a uniform distribution and the distribution of the mapping to target classes. R-FGSM and P-FGSM are the closest to a uniform distribution with a distance of 0.1414 and 0.1416, respectively. When the distribution is closer to a uniform distribution, the deduction of the true class of a transformed image is more difficult.

Figure 3 shows how often each of the 365 classes is selected as target class, regardless of the true class of each of the 3,000 images of the Places365-Standard dataset. R-FGSM and P-FGSM have a similar distribution, which is close to uniform and thus desirable. L-FGSM has a significantly less uniform distribution and, for example, the same target class is selected 6.77% of the times.

In summary, while P-FGSM and R-FGSM are comparable in terms of *irreversibility*, P-FGSM has a higher misclassification rate than R-FGSM, thus indicating better performance.

4. CONCLUSION

In this paper, we discussed how adversarial images can be exploited for privacy protection against automatic inference and proposed *private* FGSM, an algorithm for privacy protection. Images transformed with *private* FGSM maintain a good visual quality and, compared to images generated with other adversarial approaches, have a higher degree of irreversibility. We showed that, in a scene privacy-protection task, *private* FGSM always misleads a ResNet50 multi-class classifier in its top-1 result and 94.40% of the times in its top-5 results.

As future work, we will evaluate the detectability of methods,

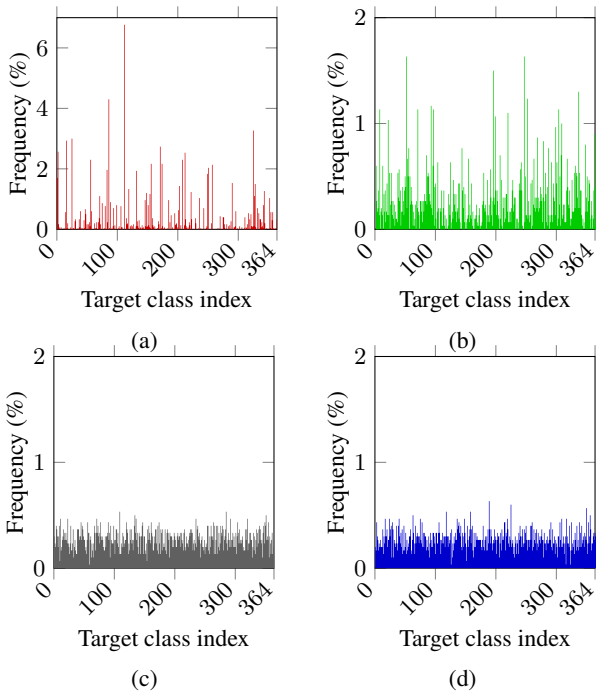

Fig. 3. Frequency of selection of each of the 365 classes of the testing subset of Places365-Standard [21] generated by (a) L-FGSM, (b) N-FGSM, (c) R-FGSM and (d) P-FGSM. The more uniform the distribution, the higher the irreversibility. Note the different scale of the vertical axis in (a).

and validate the proposed approach on other datasets and with other classifiers, as well as explore its applicability for privacy protection in other classification tasks.

5. REFERENCES

[1] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in Neural Information Processing Systems (NIPS)*, Montreal, Canada, December 2014.

[2] A. Senior, S. Pankanti, A. Hampapur, L. Brown, Y.-L. Tian, A. Ekin, J. Connell, C.F. Shu, and M. Lu, "Enabling video privacy through computer vision," in *IEEE Security Privacy*, Oakland, California, USA, May 2005.

[3] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "Cartoongan: Generative adversarial networks for photo cartoonization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, USA, June 2018.

[4] K. Chinomi, N. Nitta, Y. Ito, and N. Babaguchi, "PriSurv: privacy protected video surveillance system using adaptive visual abstraction," in *International Conference on MultiMedia Modeling (MMM)*, Kyoto, Japan, January 2008.

[5] O. Sarwar, B. Rinner, and A. Cavallaro, "Design space exploration for adaptive privacy protection in airborne images," in *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Colorado Springs, CO, USA, August 2016.

[6] O. Sarwar, A. Cavallaro, and B. Rinner, "Temporally smooth privacy protected airborne videos," in *IEEE/RSJ International*

Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, October 2018.

[7] S. Ciftçi, A. O. Akyüz, and T. Ebrahimi, "A reliable and reversible image privacy protection based on false colors," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 68–81, January 2018.

[8] F. Dufaux and T. Ebrahimi, "Scrambling for privacy protection in video surveillance systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1168–1174, August 2008.

[9] P. Korshunov and T. Ebrahimi, "Using warping for privacy protection in video surveillance," in *International Conference on Digital Signal Processing (DSP)*, Santorini, Greece, July 2013.

[10] J. Chen, J. Konrad, and P. Ishwar, "VGAN-based image representation learning for privacy-preserving facial expression recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, USA, June 2018.

[11] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, USA, June 2016.

[12] Y. Liu, W. Zhang, and N. Yu, "Protecting privacy in shared photos via adversarial examples based stealth," *Security and Communication Networks*, vol. 2017, Art. ID. 1897438, 15 pages, November 2017.

[13] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, USA, July 2017.

[14] J. Su, D. Vasconcellos Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *CoRR*, vol. abs/1710.08864, 2017.

[15] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, San Diego, USA, 2015.

[16] D. Warde-Farley and I. Goodfellow, "Adversarial perturbations of deep neural networks," in *Perturbations, Optimization, and Statistics*, T. Hazan, G. Papandreou, and D. Tarlow, Eds., chapter 11, pp. 311–343. The MIT Press, Cambridge, Massachusetts, London, England, 2017.

[17] A. Kurakin, I. Goodfellow, S. Bengio, Y. Dong, F. Liao, M. Liang, T. Pang, J. Zhu, X. Hu, C. Xie, J. Wang, Z. Zhang, Z. Ren, A. Yuille, S. Huang, Y. Zhao, Y. Zhao, Z. Han, J. Long, Y. Berdibekov, T. Akiba, S. Tokui, and A. Abe, "Adversarial attacks and defences competition," in *arXiv:1804.00097*, 2018.

[18] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *International Conference on Learning Representations (ICLR) Workshop Track*, Toulon, France, April 2017.

[19] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *International Conference on Learning Representations (ICLR)*, Toulon, France, April 2017.

[20] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, June 2018.

[21] "Pixel privacy task, mediaeval 2018," http://www.multimediaeval.org/mediaeval2018/, 2018, [Last accessed February 2019].

[22] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[23] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, December 2012.