Adversarial Mask Generation for Preserving Visual Privacy

Aayush Gupta, Ayush Jaiswal, Yue Wu, Vivek Yadav, and Pradeep Natarajan
Amazon Alexa AI Natural Understanding

***Abstract*— We present a privacy preserving machine learning method for images that separates task-relevant information from task-irrelevant information. Our primary hypothesis is that by revealing the minimal number of pixels required for a task we can provide the most privacy preserving guarantees. Specifically, we propose an adversarial method that masks out task-irrelevant information from an image for preserving privacy. The proposed method only uses task-specific label information and no privacy annotations such as identity of the subject, gender, race, *etc.*, are required. We validate the proposed method on face attribute prediction on the CelebA dataset and emotion recognition on the FER+ dataset, showing that we can preserve visual privacy with little degradation in the task performance.**

I. Introduction and Related Work

Machine learning (ML) has become increasingly important to a growing number of user-facing applications, from navigation and trip planning to recommendations for which movies to watch. Concomitant with the increasing use of ML, there is a growing interest in privacy preserving approaches to learning that aim to keep both the data and the ML models secure. While cybersecurity measures can be adopted to safeguard the stored data, the private information implicitly recorded within ML models requires more sophisticated approaches [8], [27], [3]. In order to better protect user-privacy in the face of malicious attacks, we aim to build machine learning methods that minimize the task irrelevant data exposed to various stakeholders like data annotators, analysts and designers of the ML models, and to the ML models themselves. Several recent works aim to provide mathematical definitions of privacy [5], [15], [24], [10] and methods for controlling it such that the privatized mechanisms still hold much of their utility (for analytics, model training, *etc.*) [8], [27], [3], [16], [1], [4], [11], [18], [28], [25], [23], [19], [17], [26].

Most methods for privacy-preservation in ML fall into one of five categories — (a) modification of the model training through incorporation of a theoretical notion of privacy [16], [1] as regularization, (b) adversarial training to reduce identifiability [11], [26], (c) usage of synthetic data with similar underlying distribution as real data [28], [29], (d) remapping of real data to templates through encoder-decoder based methods [25], [23], and (e) automatic redaction or obfuscation of sensitive parts of data [19], [17]. While approaches (a) and (b) focus on changing the model training for privacy-preservation, methods (c)–(e) aim to modify the data itself such that all identifying information is removed before usage for model training.

This work was not supported by any organization.

978-1-6654-3176-7/21/$31.00 ©2021 IEEE

The training-focused approaches to privacy-preservation incorporate theoretical models of privacy as regularizers into model training. For example, Abadi *et al.* [1] propose a modification of the stochastic gradient descent optimization method for training NNs to conform to Differential Privacy [8]. The work of Nasr *et al.* [16], on the other hand, proposes a regularization for promoting Membership Privacy [10] in ML models. Besides methods that involve theoretical definitions of privacy, adversarial training schemes have been proposed to reduce the amount of protected information encoded by NNs. For example, Li *et al.* [11] train NNs against a discriminator for text-based tasks such that specific protected attributes like age, gender, *etc.*, are not encoded, whereas the training scheme of Wu *et al.* [26] involves adversarially learning a degradation transform that reduces the information seen by the NN on-the-fly.

In the data-focused class of methods, works like [28], [29] have proposed methods for training Generative Adversarial Networks (GANs) [6], a class of deep generative models, with differential privacy such that "privatized" synthetic data can be generated from such GANs where the synthetic data is expected to have a similar distribution as the real data. Methods have also been proposed to transform data samples by mapping them to templates such that only information relevant for the end-use is reflected in the transformed version and all other information is hidden. For example, Wu *et al.* [25] and Shirai *et al.* [23] both propose GAN-based approaches for transforming face images for privacy-preservation. Similarly, Maximov *et al.* [14] propose a conditional GAN to anonymize face and body images/videos.

Automatic redaction or obfuscation methods have also been proposed to remove sensitive parts of data samples in order to protect privacy. Raval *et al.* [19] use adversarial training to perturb images using an obfuscator network such that the produced image does not contain specific protected information but is similar to the original image for maximizing downstream utility. Orekondy *et al.* [17], on the other hand, present a method for detecting sensitive text, faces, persons, *etc.*, in images for subsequent redaction.

In this work, we focus on the problem of automatic redaction of visual data (category (e)) such that only necessary parts of an image sample are revealed for specific annotation, analysis, and modeling purposes. As such, the revealed content is maximally informative for the particular task but minimally so overall, *e.g.*, a person's background could be masked for face identification, only the hair of an individual could be revealed for analyzing hair type (straight or curly), *etc.* While it is possible to achieve this through semantic segmentation [9] by training models with

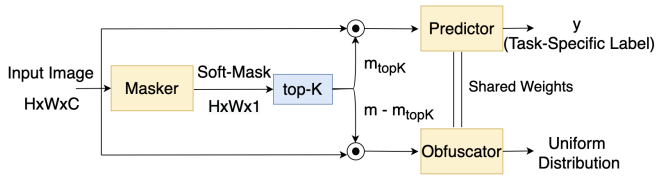

Fig. 1. Model Architecture: An input image is passed to a Masker network that produces a soft-mask of the same resolution. Next, top-K% pixels are chosen from this soft-mask(rest of the pixels are zeroed out). An elementwise product between the soft mask, $m_{\text{topK}\%}$, and the image is passed to the Predictor network and the complementary mask, $m - m_{\text{topK}\%}$, with the image to the Obfuscator network. The model is trained adversarially such that the input to the predictor captures all task-relevant information.

pixel-level annotations of what different parts of an image indicate, such annotations are expensive and training models for semantic segmentation is not easy as it requires large amounts of annotated data and complex machine learning models [9]. In contrast, we propose a model that can generate occlusion masks using simple annotations that are indicative of the portions of the sample to be retained, *e.g.*, subject identity labels for retaining faces while masking out the background, hair-type annotation for retaining only the hair in head images, *etc.*

The proposed model learns to produce pixel-level binary masks that split the image into two components – one that contains only the parts of the image deemed necessary for the specific task and the other that contains everything else. Thus, the proposed approach is expected to be conservative in revealing information, making the resulting images highly privatized. Our hypothesis is that for a particular task, the entire image is not required to make a good prediction. For instance, in order to recognize the identity of a person, background pixels are not required.

The model works by first feeding the input images x through a Masker module, which is a neural network (NN) that generates a soft pixel-level mask m. We then select top-K% pixels from the generated mask to get another mask $m_{\text{topK}\%}$ (top K% pixels are kept with there soft scores and rest of the pixels are zeroed out) that is used to split the image into the two components, $x_p = m_{\text{topK}\%} \odot x$ and $x_o = (m - m_{\text{topK}\%}) \odot x$, where $\odot$ denotes elementwise multiplication. Next, a Predictor NN is used to infer the simple annotation y from both x_p and x_o. The complete model is trained adversarially such that at convergence, the predictor should be able to predict y from x_p but not from x_o. Hence, x_p and x_o correspond to the revealed (task-relevant) and the protected components (task-irrelevant) of the image, respectively, and x_p can be released for the end-use, *e.g.*, training downstream machine learning models, data-annotation tasks, *etc.* Experimental evaluation of the proposed model is conducted on two datasets – CelebA [12] and FER+[2]. Results show that the proposed model is effective at identifying the task-specific parts of the image and masking out everything else, such that the revealed components exhibit high utility scores on the final use-case and low identifiability scores for sensitive information.

II. Learning to Mask

A. Model Architecture

Figure 1 presents the model architecture for our privacy-preserving framework. We first pass an image x through a Masker NN that generates a task-specific mask for the image. We use a standard U-Net [20] like architecture, which is widely used in segmentation literature, to generate pixel-level masks $m = \text{UNet}(x)$. We next select top-K% of the pixels from m, treating them as pixels of interest. The top-K% mask pixels are selected with their soft value and the others are zeroed out during training. The parameter K gives us precise control of the percentage of pixels revealed to the prediction model. For example, an ML practitioner can constrain the prediction model to use only 25% of the pixels to predict hair color and ignore the other pixels. The resulting mask $m_{\text{topK}\%}$ is then used to split the image into $x_p = m_{\text{topK}\%} \odot x$ and $x_o = (m - m_{\text{topK}\%}) \odot x$, where x_p and x_o are the inputs to the Predictor and the Obfuscator respectively. The Predictor and the Obfuscator share the same weights and architecture, and we use Φ to refer to the shared underlying model. The predictions from the Predictor and the Obfuscator are denoted as $y_p = \Phi(x_p)$ and $y_o = \Phi(x_o)$, respectively.

B. Model Training

We train our model in an adversarial manner, where in the main step, the Predictor and the Masker are updated to use top K% of the pixels to give the best performance, whereas in the adversarial step, the Masker is updated so that the Obfuscator is the most confused, i.e., predictions are as close to random as possible. We apply three batch gradient updates during the Predictor step, and one batch for the Obfuscator step. During training, we train to maximize prediction accuracy from the Predictor and obfuscate/confuse the predictions from the Obfuscator as much as possible. In order to achieve this, we train the Predictor to minimize the cross entropy loss for the task-specific labels, similar to any other classification method: $L_{\text{Predictor}} = \sum_i^N -y_i * log(y_{p_i})$, where y_i is the task label and y_{p_i} is the prediction from the Predictor. In this step, we optimize the weights of the Predictor NN and the Masker (UNet) network to minimize the cross entropy loss. In order to confuse the Obfuscator model, we minimize the cross entropy between the prediction and uniform distribution: $L_{\text{Obfuscator}} = \sum_i^N -1/N * log(y_{o_i})$, where N is the number of classes in the classification task and y_{o_i} is the prediction from the Obfuscator. In this step, we only optimize the weights of the Masker (UNet) network to minimize the Obfuscator loss so that the complementary mask results in maximum confusion for the model.

III. Experimental Evaluation

A. Datasets

CelebA dataset for Facial Attribute Identification: We test our method on the task of facial attribute identification. We learn to mask the parts of a person's face that are not relevant to the task of facial attribute identification. We choose hair color as the facial attribute to be predicted. The

TABLE I

RESULTS FOR FACIAL EMOTION RECOGNITION. WE REPORT THE MEAN FACE SIMILARITY AS A METRIC FOR PRIVACY PRESERVATION AND FER ACCURACY AS A METRIC FOR TASK-UTILITY PERFORMANCE.

Method	Face Similarity	FER
Original Image	1.00	92.49
Block	0.16	81.34
Gaussian Blur	0.35	87.78
CIAGAN + Dlib Face & Shape Predictor	0.19	87.79
DeepPrivacy + DSFD + Mask-RCNN	0.26	88.01
Our Method – Non-Adversarial	0.48	89.81
Our Method	0.20	88.64

TABLE II

RESULTS FOR FACIAL HAIR COLOR IDENTIFICATION. WE REPORT THE MEAN FACE SIMILARITY AS A METRIC FOR PRIVACY PRESERVATION AND HAIR COLOR ACCURACY FOR TASK-UTILITY PERFORMANCE.

Method	Face Similarity	Hair Color
Original Image	1.00	89.01
Block	0.36	78.89
Gaussian Blur	0.72	85.86
CIAGAN + Dlib Face & Shape Predictor	0.31	85.57
DeepPrivacy + DSFD + Mask-RCNN	0.49	84.49
Our Method – Non-Adversarial	0.47	86.73
Our Method	0.35	85.59

privacy task is defined as identifying the subject identity. We employ the widely used CelebA [12] dataset with the standard train/valid/test split. We only use the images in these splits for which hair color attribute is available. Hence, we have 74,024 images for training, 18,505 images for validation and 11,351 images for testing.

FER+ dataset for Facial Emotion Classification: We additionally test our framework on the facial emotion classification task. In the FER+ dataset [2] each image has multiple emotion tags from 10 annotators. We extract the majority vote emotion associated with each image and frame the problem as multi-class classification over the emotion labels. Further, we select only the images with the three dominant emotion classes namely neutral, happiness and surprise due to lack of data for the other classes. The privacy task is defined as identifying the subject identity. We use the standard dataset split for this task, with 15,616 images for training, 1,943 for validation and 1,930 for testing.

B. Experimental Setup

Implementation details: We use TensorFlow for the implementation of our framework. The masker has a U-Net architecture as described previously, which progressively downsamples the input image till a bottleneck layer and then upsamples the hidden information to reconstruct an image of the same resolution but with only a single channel (64x64x1). In the final layer, we use sigmoid activation to produce a soft mask. The soft mask is then thresholded to select only the top-K% pixels and rest are discarded/zeroed-out.The Predictor and the Obfuscator are classification networks with four convolutional layers (with batch-normalization, ReLU and max-pooling (2x2 filter size)), one dense layer (followed by batch-normalization and ReLU) and the final classification dense layer with soft-max activation.

Baselines: We use three baselines for our experiments. The first baseline is the original image blurred by a Gaussian kernel at the center (Gaussian Blur). Since the images are centered and aligned, this hides most of the face. In the second baseline, we black out the center part of the image, completely hiding the face of the person (Block). The third baseline is our model (Masker and Predictor) without the Obfuscator trained in a non-adversarial manner. The results for this baseline are reported with the same top-K% hyperparameter as our model trained adversarially. We also compare our method against two recent face anonymization methods namely CIAGAN [14] and DeepPrivacy [7].

Evaluation: In order to evaluate the efficacy of our method quantitatively, we show the downstream utility task performance for our masked images. Specifically, we show that the masked images produced by our method maintain high Facial Hair Color Identification accuracy for the CelebA dataset and high Facial Emotion Recognition accuracy for the FER+ dataset. To measure Facial Hair Color accuracy we use SlimCNN[22] and for facial emotion recognition we used FERAtt [13]. Furthermore, the privacy preservation aspect is shown using perceptual similarity between the masked and original images, namely Face Similarity. It is the cosine similarity between the deep features of the images using a pretrained model. The pretrained model used is FaceNet[21].

C. Results and Analysis

Facial Hair Color Prediction: In Table 1, we show empirical results for our method for the task of hair color prediction on the CelebA dataset. Our method maintains good hair color prediction performance as compared to just using the original image. Furthermore, it outperforms the Block baseline in hair color prediction performance while maintaining more privacy with less face-similarity. For the Gaussian Blur baseline, the model has similar face hair color prediction accuracy but our model preserves the privacy far better with much lower face similarity. In comparison to the non-adversarial baseline we have better privacy protection but less hair color prediction accuracy. This is to be expected since the non-adversarial model only optimizes on the task performance and does not necessarily discard pixels irrelevant to the task performance. Our method, outperforms the DeepPrivacy baseline in both prediction accuracy and face similarity. As compared to CIAGAN, our method performs comparably for both task performance and privacy preservation. However, CIAGAN uses face detection and face landmark detection as a preprocessing step whereas our method uses simple task-labels (hair color), making our model both simpler to train and more lightweight to use. In Figure 2, we present visualizations of masked images for the baselines and our model, showing that our model reveals high-utility meaningful image parts while hiding non-essential subject-identifiable parts.

Facial Emotion Recognition: In Table 2, we show empirical results for our method for the task of facial emotion recognition on the FER+ dataset. In these experiments, our model

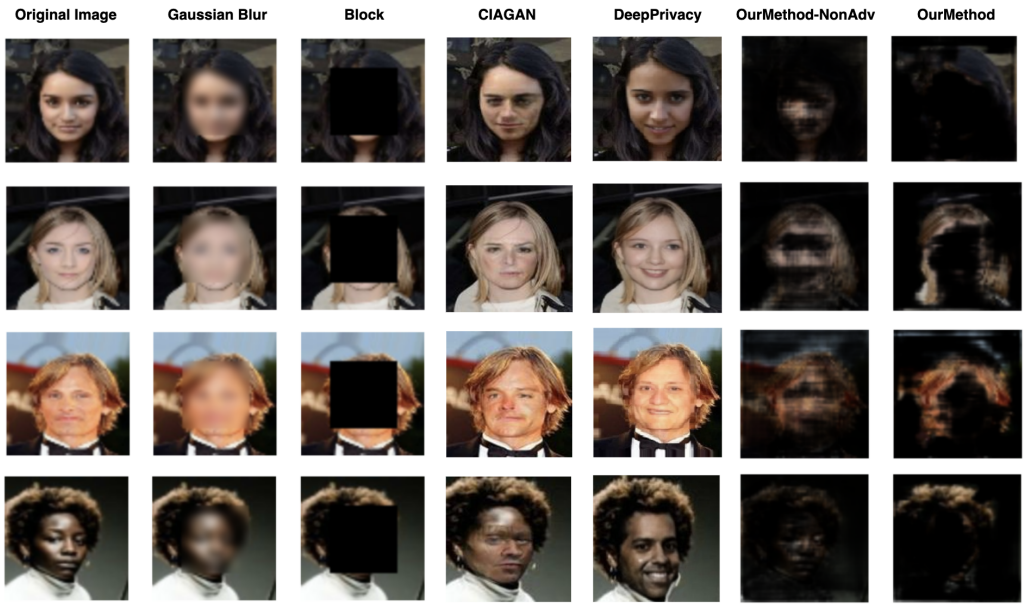

Fig. 2. Visualization for hair color identification. Original Image is on the extreme left, followed by Gaussian Blur, Block, CIAGAN, DeepPrivacy and the Non-adversarial baseline. Masked images from our method are on the extreme right.

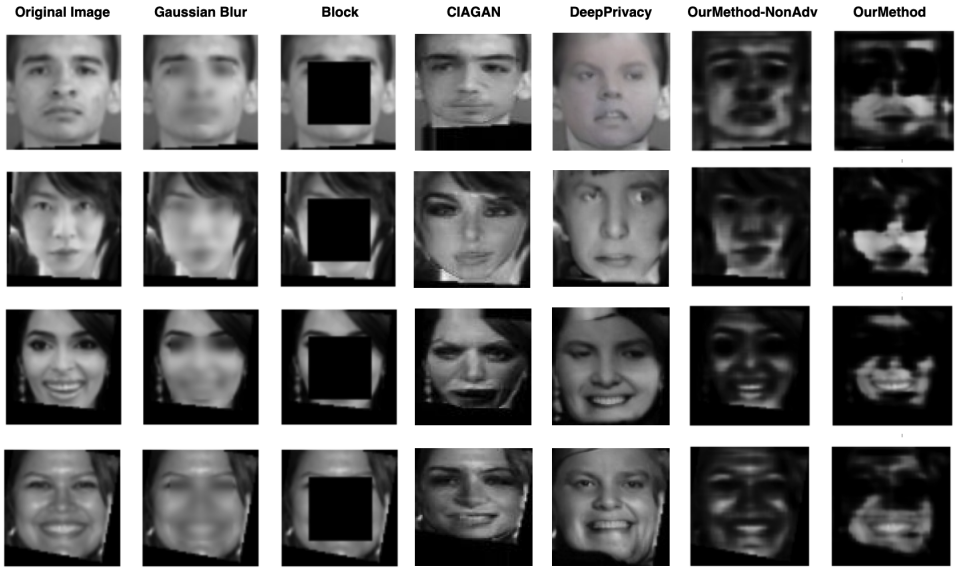

Fig. 3. Visualization for Facial Emotion Recognition. Original Image is on the extreme left, followed by Gaussian Blur, Block, CIAGAN, DeepPrivacy and the Non-adversarial baseline. Masked images from our method are on the extreme right.

outperforms the Gaussian Blur baseline on both the facial emotion prediction performance and face similarity. In the case of Block baseline, although the privacy preservation is worse for our model, it performs significantly better on the facial emotion prediction task. In comparison to the non-adversarial baseline, the adversarial model preserves privacy better but at the cost of some task performance. Our method outperforms DeepPrivacy method and performs comparably to CIAGAN similar to the hair color prediction task but maintains the inherent advantage of using only the task level labels (facial emotion). Figure 3 presents the visualization of masked images for the baselines and our model, showing that our model reveals only parts of the image that are relevant to facial emotion recognition.

IV. Conclusion

We have proposed a novel end-to-end framework of preserving visual privacy that masks out the information that is not required for a downstream machine learning task or a data-annotation task. Furthermore, the method not only separates task-critical information from private information but also produces a mask that localizes this information. Unlike previous methods, our model only requires task-relevant macro-annotations (*e.g.*, hair color) and automatically learns to produce this separation of image parts. Experimental results on two datasets show that our method is able to preserve privacy as well as maintain task-utility performance.

REFERENCES

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 308–318, New York, NY, USA, 2016. Association for Computing Machinery.

[2] E. Barsoum, C. Zhang, C. Canton Ferrer, and Z. Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ACM International Conference on Multimodal Interaction (ICMI)*, 2016.

[3] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, Santa Clara, CA, Aug. 2019. USENIX Association.

[4] A. T. Chen, M. Biglari-Abhari, and K. I. Wang. Trusting the computer in computer vision: A privacy-affirming framework. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1360–1367, 2017.

[5] P. Cuff and L. Yu. Differential privacy as a mutual information constraint. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 43–54, New York, NY, USA, 2016. Association for Computing Machinery.

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

[7] H. Hukkelås, R. Mester, and F. Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *Advances in Visual Computing*, pages 565–578. Springer International Publishing, 2019.

[8] Z. Ji, Z. C. Lipton, and C. Elkan. Differential privacy and machine learning: a survey and review, 2014.

[9] F. Lateef and Y. Ruichek. Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, 338:321 – 348, 2019.

[10] N. Li, W. Qardaji, D. Su, Y. Wu, and W. Yang. Membership privacy: A unifying framework for privacy definitions. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security*, CCS '13, page 889–900, New York, NY, USA, 2013. Association for Computing Machinery.

[11] Y. Li, T. Baldwin, and T. Cohn. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[12] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[13] P. D. Marrero Fernandez, F. A. Guerrero Pena, T. Ing Ren, and A. Cunha. Feratt: Facial expression recognition with attention net. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[14] M. Maximov, I. Elezi, and L. Leal-Taixe. Ciagan: Conditional identity anonymization generative adversarial networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020.

[15] D. J. Mir. Information-theoretic foundations of differential privacy. In J. Garcia-Alfaro, F. Cuppens, N. Cuppens-Boulahia, A. Miri, and N. Tawbi, editors, *Foundations and Practice of Security*, pages 374–381, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[16] M. Nasr, R. Shokri, and A. Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, page 634–646, New York, NY, USA, 2018. Association for Computing Machinery.

[17] T. Orekondy, M. Fritz, and B. Schiele. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[18] J. Qian, F. Han, J. Hou, C. Zhang, Y. Wang, and X. Li. Towards privacy-preserving speech data publishing. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pages 1079–1087, 2018.

[19] N. Raval, A. Machanavajjhala, and L. P. Cox. Protecting visual secrets using adversarial nets. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1329–1332, July 2017.

[20] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[21] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[22] A. K. Sharma and H. Foroosh. Slim-cnn: A light-weight cnn for face attribute prediction. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 329–335, 2020.

[23] S. Shirai and J. Whitehill. Privacy-preserving annotation of face images through attribute-preserving face synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[24] W. Wang, L. Ying, and J. Zhang. On the relation between identifiability, differential privacy, and mutual-information privacy. *IEEE Transactions on Information Theory*, 62(9):5018–5029, 2016.

[25] Y. Wu, F. Yang, Y. Xu, and H. Ling. Privacy-protective-gan for privacy preserving face de-identification. *Journal of Computer Science and Technology*, 34(1):47, 2019.

[26] Z. Wu, Z. Wang, Z. Wang, and H. Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 627–645, Cham, 2018. Springer International Publishing.

[27] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. *IEEE Computer Security Foundations Symposium*, Jul 2018.

[28] J. Yoon, J. Jordon, and M. van der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2019.

[29] X. Zhang, S. Ji, and T. Wang. Differentially private releasing via deep generative model (technical report), 2018.