2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Visual Coherence Face Anonymization Algorithm Based on Dynamic Identity Perception

Xuan Tan[1], Shanqing Zhang[1], Yixuan Ju[2], Xiaoyang Mao[2], Jiayi Xu[1*]
[1] School of Computer Science, Hangzhou Dianzi University, China
[2] Department of Computer Science and Engineering, University of Yamanashi, Japan

***Abstract*— In the era of the meta-universe and the proliferation of personalized social networks, interactive behaviors like sharing personal and family photos pose an escalating risk of privacy breaches and identity exposure. A potential remedy lies in substituting real images with anonymized face images in public contexts. While existing face anonymization methods often replace substantial portions of face images, the resultant faces lack sufficient similarity to the originals. To address this, we propose a anonymization model leveraging saliency analysis to detect identity relevant facial region, preserving visual coherence and avoiding recognition by face recognition systems. Our model comprises two integral networks: the Dynamic Identity Perception Network (DIPNet) and the improved PSPNet. DIPNet, in particular, encompasses two vital sub-modules: the dynamic region perception module detects identity relevant region; the anonymization region control module governs the size of region through thresholding, thereby dominating the preservation of identity independent features and the degree of anonymization. The improved PSPNet produces high-quality identity anonymized faces. Experimental results demonstrate that our method yields realistic anonymized faces, retaining original features and deceiving face recognition systems, safeguarding privacy in the modern digital landscape.**

I. INTRODUCTION

The exponential growth of social networks has integrated photo and video sharing seamlessly into people's digital entertainment experiences. While the collection and sharing of personal information enhance convenience and happiness, the flourishing digital economy raises significant concerns about facial privacy [25]. Face anonymization techniques, aimed at safeguarding individuals' privacy in images or videos by concealing or eliminating recognizable identities, have garnered increasing attention in response to these growing concerns [3].

Early face anonymization techniques [17], [19], including blurring and pixelation, often alter faces drastically, compromising image quality and usability. The K-Same algorithm [18] addresses this by replacing similar faces with an average face, yielding k identical anonymized faces and reducing identity matching probability to 1/k. However, K-Same masks crucial individual details, potentially losing attributes like race and facial expressions. Recent deep learning based methods involve latent code modification methods [11], [14], [28], manipulating latent codes to disable face recognition, but resulting in significant differences from the original face due to attribute entanglement. Facial region modification methods [7], [10], [15] erase and inpaint large face region, neglecting control over filled region features, leading to uncontrollable attribute generation. These challenges underscore the need for advanced face anonymization techniques that balance identity protection, attribute control, and downstream task compatibility.

(a) Salient region for masked faces with the same identity.

(b) Salient region for masked faces with the different identities.

Fig. 1. Example of masked faces with same and distinct identities from VggFace2 dataset.

In order to minimize face modification while preventing being identified by face recognition system, we propose a face anonymization model based on the dynamic perception of identity salient region to minimize the modification area. The proposed model comprises two integral networks: the Dynamic Identity Perception Network (DIPNet) and the improved PSPNet [20]. DIPNet consists of two vital sub-modules: the dynamic region perception module is able to obtain the core face region related to the identity based on saliency analysis, and generate a ID attention map on the face for modification constraint purpose; the anonymization region control module generates a mask through thresholding. By multiplying it with the original image, a masked face is obtained, maximizing the preservation of identity independent features. Finally, through an improved end-to-end PSPNet with four losses, we obtain anonymized images,

This research was funded by the National Natural Science Foundation of China (grant No.62102125 and No.62172132).

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

striking a balance between anonymization and similarity.

- We propose face anonymization method to minimizes modification by focusing on the dynamic perception of identity.
- We obtain identity-related masks through threshold segmentation, leaving a large untouched area for visual coherence with the original face. This strategy ensures effective face de-identification while preserving facial features.
- Four loss components harmonize privacy and similarity throughout the image generation process, encompassing de-identification loss, regularization loss, reconstruction loss, and perceptual loss.

II. RELATED WORK

a) Face editing: Generative Adversarial Network (GAN [4]), synthesizing high quality virtual face images becomes achievable. To overcome random face generation limitations, Conditional GAN [16] have been introduced. By incorporating a conditional vector onto the noise vector, these models offer control over face attribute changes. The disentanglement of latent codes facilitates intuitive control over various face attributes in the GAN latent space. InterFaceGAN model [24] back projects the faces to get corresponding latent codes, calculates attribute score for each latent code in the dataset, then uses Support Vector Machine (SVM) to solve the hyperplane of 5 typical attributes in latent space to obtain attribute control vectors. Xu et al. [26] propose the TransEditor model which introduced a cross-space attention mechanism based on Transformer, and can flexibly control the level of modification along different attribute directions to achieve flexible face editing.

b) Face anonymization: Researchers have explored face image editing as a means to deceive face recognition systems and safeguard identity. Traditional methods like face mosaics or gaussian blur compromise image quality [19]. K-Same based techniques [18] replace k original faces with one anonymized face, risking a loss of face diversity.

GAN based models have emerged to reform face images, categorized into latent code modification methods and facial region modification methods. Latent code modification methods modify face images in latent space, erasing identity features, yet decoder regeneration may lose facial features. Ma et al. [14] projected the face into ID representation space, selected the face that is far away from the original face in ID feature for face swap, and achieved the purpose of controlling the degree of face anonymization. Zhai et al. [28] treated face de-identification as a joint task of semantic suppression and controllable attribute injection. The semantic suppression network removed the identity-sensitive information, while the attribute-aware injective network generated DeId-sensitive attributes in a controllable way. On the other hand, facial region modification methods focus on modifying facial attributes to indirectly change the identification information, or removing a part from the face region and then filling. Maximov et al. [15] proposed CIAGAN model to anonymize face images and videos based on CGAN. The landmarks were first extracted to get face contour information, a mask covering the contour was obtained and filled with faces of other identities in the dataset. Kuang et al. [10] proposed DeIdGAN model that uses semantic segmentation technique to obtain face mask region to be modified, and fill the mask region with face images of other identities mixed with random noise.

Overall, current face anonymization research aims to conceal identity while preserving face similarity. Latent code modification methods often impact other attributes, leading to noticeable differences. Facial region modification methods, as large part of the face is filled randomly or swapped with face of other identity, risk losing key attributes, affecting downstream tasks like facial expression recognition.

III. PROPOSED METHOD

The core objective of face anonymization is to protect privacy through the obscuration of identity information. However, the balance between high unrecognizable rate and minor facial modification remains to be a technical challenge. To address this, we propose a model comprises two integral networks: the Dynamic Identity Perception Network (DIPNet) and the improved PSPNet. DIPNet, in particular, encompasses two vital sub-modules: the dynamic region perception module identifies salient identity region, associating identity with specific facial region; the anonymization region control module employs threshold segmentation to isolate a mask, enabling focused anonymization edits on a confined region. Finally, the improved PSPNet extracts identity-independent features from the masked face, ensuring high visual consistency with the original faces.

A. Dynamic identity-related region perception

We design dynamic region perception module to adaptively allocate the core face region which associated with the identity. Later on, we can perform image inpainting within the obscured region to synthesize face with obscured identity. The generation of ID attention map consists of two steps as follows:

a) Face cropping: To detect the face in an image, we use Dlib [12] to allocate 68 key feature points. We crop the face region including the area from the upper eyebrow to the chin because this part contains majority of the identity features with an identification accuracy of 99% in our experiments. Additionally, constrain the following facial modification within the crop region retains original features and guarantee less visual differences.

b) Identity salient region extraction: we appoint unique identity in the form of one-hot vector for all faces in the VggFace2 [1] and CelebA [13] datasets. To determine the face region, which is most associated with identity [27], we employ Grad-CAM [29], [22] to get a weighted heat map that indicates the importance prediction of the input image. More specifically, the cropped face image is input into a trained SE-ResNet50 [6] face recognition network to predict its identity classification probability. Then, the

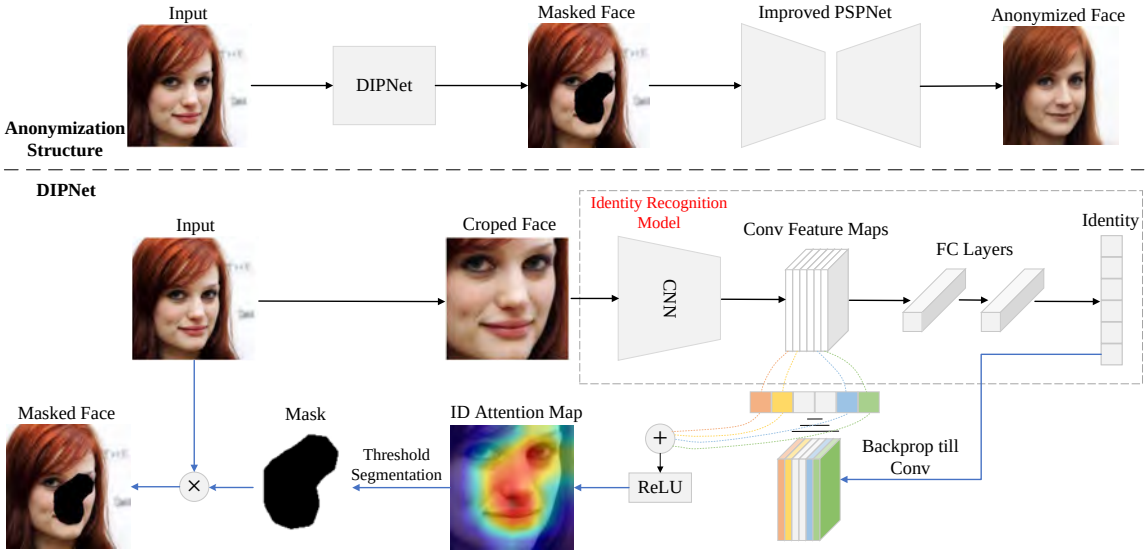

Fig. 2. Architecture of the proposed anonymization model and the Dynamic Identity Perception Network (DIPNet). For DIPNet, given an input face image, identity-related region is firstly detected through the dynamic region perception module and an ID attention map is generated. Then, utilizing the anonymization region control module with threshold segmentation, we identify the region mask associated with identity, ensuring the optimal preservation of identity-independent features. The multiplication of the mask with the original image produces the masked face. Finally, the masked face is input into the improved PSPNet to obtain an anonymized face.

gradient of the identity classification probability together with the last feature layers of SE-ResNet50, are combined to get a weighted ID attention map. As shown in Fig. 2, the area with higher value (indicated by red in the pseudo-colored ID attention map) corresponds to a greater contribution to the identity compared with the blue area.

B. Anonymization region control

We segment the ID attention map and apply binarization using a threshold. Pixels above the threshold are set to 1, and those below are set to 0. A higher threshold value means a smaller modification area. The resulting masked face is generated by multiplying the original face with the binarized ID attention map. By adjusting the threshold value, we can control the extent to which identity-related region is masked, thereby influencing the degree of anonymity and suit specific privacy and data reusability requirements.

C. Improved PSPNet

The original PSPNet [20] employs a standard feature pyramid over the ResNet backbone to extract 512-dimensional vectors from feature maps. These vectors are then input into a pre-trained StyleGAN generator [8], [9] for facial image generation. Incorporating the dynamic region perception module and anonymization control module, DIPNet identifies salient facial region, yielding a masked face. However, for optimal balance between privacy preservation and visual coherence, the mask coverage is constrained in certain identity-related region. The original PSPNet, might still deduce identity by using masked face, leading to less effective anonymization. To address this, we improved the original PSPNet specifically for anonymization purposes by introducing a de-identification loss during joint training, with the goal of maximizing the elimination of identity.

$$L_{deid}(X) = cos(R(A_f(X)),\ R(X)) \tag{1}$$

where $A_f(\cdot)$ denotes the anonymization model, $R(\cdot)$ is the pretrained ArcFace face recognition network [2]. $cos(\cdot,\cdot)$ denotes cosine similarity. By incorporating the de-identification loss, we strive to enhance the effectiveness of our proposed anonymization process and minimize the potential leakage of identity in the non-concealed region.

The remaining three losses are consistent with those in PSPNet, among which the regularization loss L_{reg} ensures quality generation of images with a blacked-out area by guiding the anonymized face to approximate the average face; reconstruction loss L_2 promotes pixel-wise similarity; perceptual loss L_{per} leverages the VGG16 network to learn perceptual similarities.

$$\begin{aligned} L_{total}(X) = \lambda_1 L_{reg}(X) + \lambda_2 L_2(X) + \lambda_3 L_{per}(X) \\ + \lambda_4 L_{deid}(X) \end{aligned} \tag{2}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are hyper-parameters. By combining these loss functions into a comprehensive framework, the carefully tuned hyper-parameters enable us to strike a balance between preserving face features and achieving the desired degree of identity anonymization.

IV. EXPERIMENTS AND ANALYSIS

To validate the efficiency of the proposed model, two experiments are conducted. We utilize the SE-ResNet50 [6] and FaceNet [21] models for identity recognition rate assessment after anonymization. FID (Fréchet Inception Distance [5]) quantifies the quality and dissimilarity between the distribution of anonymized faces and real faces post-anonymization, and the Attributes Rate evaluates attribute preservation after anonymization based on ResNet50.

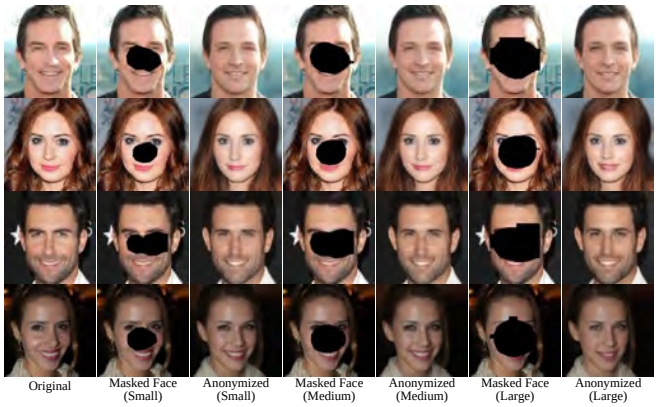

Fig. 3. Visual comparison of our results using mask of various sizes.

TABLE I

EVALUATION RESULTS ON CELEBA USING MASK OF VARIOUS SIZES.

Sizes	Recognition Rate↓		FID↓	Attributes Rate↑
	SE-ResNet50	FaceNet		
Small	2.56%	2.72%	18.79	84.82%
Medium	2.41%	2.49%	18.72	84.81%
Large	1.60%	1.64%	19.58	84.49%

A. Evaluation of the generated masked faces

In our model, DIPNet dynamically obtain the core face region related to the identity based on saliency analysis, and generate a mask on the face for modification constraint purpose. The mask size can be adjusted for varying anonymization degree. Fig. 3 demonstrates the synthesis of anonymized faces with different mask sizes. Large mask retain fewer original face features but can still maintain image quality. Table. I shows identity recognition rate, FID, and attribute recognition rate for different mask sizes. Small mask yields optimal attribute preservation but sacrifices anonymization rate, while large mask, with the best anonymization rate, results in the poorest FID and attributes preservation. For a balance between anonymization and feature similarity, a medium mask size is recommended.

We conducted another experiment on CelebA [13] and VggFace2 [1] datasets to evaluate DIPNet's performance in extracting the identity of the salient region. Fig. 1 illustrates diverse masks for faces of different identities, indicating DIPNet's ability to generate personalized and varied masks. Additionally, we measured DIPNet's effectiveness on different face images of the same identity, observing a consistent mask despite varying poses. This implies stable salient region extraction results.

B. Comparison with State-of-the-art methods

To evaluate the effectiveness of the proposed model, we compare our method with state-of-the-art (SOTA) methods, as depicted in Fig. 4. Our method yields visually pleasing and clear results. The DeIdGAN model [10] generates high-quality faces. However, its anonymization relies on attribute changes, hindering feature retention and applicability to downstream tasks. The CIAGAN model [15] assigns a new identity but suffers from imperfect face fusion, resulting in evident artifacts. Fawkes [23] exhibits visual artifacts due to added perturbation and a noise mask for identity concealment.

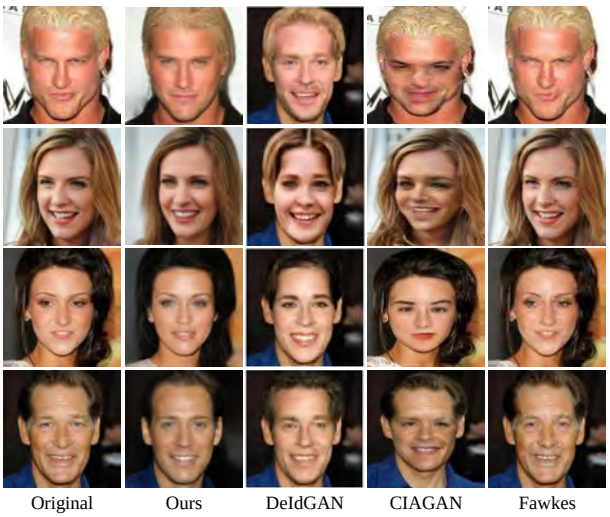

Fig. 4. Visual comparison of our results with the contrast methods on CelebA.

TABLE II

EVALUATION RESULTS ON CELEBA DATASET, WHERE ORIGINAL DENOTES THE BASELINE RESULTS USING THE ORIGINAL FACE IMAGE.

Algorithms	Recognition Rate↓		FID↓	Attributes Rate↑
	SE-ResNet50	FaceNet		
Original	99.11%	98.09%	-	87.64%
DeIdGAN	-	-	20.81	-
CIAGAN	0.89%	0.55%	34.13	81.60%
Fawkes	6.60%	5.40%	33.34	85.83%
Ours	2.41%	2.49%	18.72	84.81%

Quantitative evaluations in Table II indicate that CIAGAN achieves a slightly higher anonymization degree but at the cost of a significant decline in image quality and attributes rate, making it less applicable in practical scenarios. In contrast, our proposed method strikes a balance between image quality and recognition rate. Additionally, our model maximally retains attributes, with an attribute recognition rate closely approximating that of the original image.

V. CONCLUSION

This paper introduces an innovative face anonymization model that integrates fine-grained control in face anonymization. Our approach detects salient region closely linked to identity information, achieving minimal face modification. Simultaneously, identity-independent features from the original face are transferred to the anonymized face, introducing subtle visual differences perceptible to the human eye. Experimental results affirm the method's efficacy in ensuring high face similarity while effectively anonymizing the face.

REFERENCES

[1] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[2] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[3] J. Dietlmeier, J. Antony, K. McGuinness, and N. E. O'Connor. How important are faces for person re-identification? In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6912–6919. IEEE, 2021.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[5] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[6] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[7] S. Hu, X. Liu, Y. Zhang, M. Li, L. Y. Zhang, H. Jin, and L. Wu. Protecting facial privacy: generating adversarial identity masks via style-robust makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15014–15023, 2022.

[8] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[9] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[10] Z. Kuang, H. Liu, J. Yu, A. Tian, L. Wang, J. Fan, and N. Babaguchi. Effective de-identification generative adversarial network for face anonymization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3182–3191, 2021.

[11] T. Li and C. Clifton. Differentially private imaging via latent space manipulation. *arXiv preprint arXiv:2103.05472*, 2021.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[13] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[14] T. Ma, D. Li, W. Wang, and J. Dong. Cfa-net: Controllable face anonymization network with identity representation manipulation. *arXiv preprint arXiv:2105.11137*, 2021.

[15] M. Maximov, I. Elezi, and L. Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5447–5456, 2020.

[16] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[17] C. Neustaedter, S. Greenberg, and M. Boyle. Blur filtration fails to preserve privacy for home-based video conferencing. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 13(1):1–36, 2006.

[18] E. M. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005.

[19] J. R. Padilla-López, A. A. Chaaraoui, and F. Flórez-Revuelta. Visual privacy protection methods: A survey. *Expert Systems with Applications*, 42(9):4177–4195, 2015.

[20] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021.

[21] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[23] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *Proceedings of the 29th USENIX Security Symposium*, 2020.

[24] Y. Shen, J. Gu, X. Tang, and B. Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020.

[25] M. Taskiran, N. Kahraman, and C. E. Erdem. Face recognition: Past, present and future (a review). *Digital Signal Processing*, 106:102809, 2020.

[26] Y. Xu, Y. Yin, L. Jiang, Q. Wu, C. Zheng, C. C. Loy, B. Dai, and W. Wu. Transeditor: transformer-based dual-space gan for highly controllable facial editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7683–7692, 2022.

[27] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

[28] L. Zhai, Q. Guo, X. Xie, L. Ma, Y. E. Wang, and Y. Liu. A3gan: Attribute-aware anonymization networks for face de-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5303–5313, 2022.

[29] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.