SMRAM: Boosting Class-specific Face Privacy Protection with Stochastic Mask Regularized Adaptive Momentum

Jing Han[1], Yuanbo Li [1], Cong Hu [1]*, Xiaojun Wu [1],
[1]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China
6233111031@stu.jiangnan.edu.cn, liyuanbo12138@163.com,
conghu@jiangnan.edu.cn, wu_xiaojun@jiangnan.edu.cn

***Abstract*—To ensure secure face privacy protection, existing class-specific adversarial methods have shown substantial performance improvement by tailoring perturbations to individual identities. However, due to high local similarity in same-identity images, these methods focus perturbations on specific regions. Though some studies use generic features, they still fail to adequately balance generic and fine-grained features, yielding limited privacy protection improvements. To address these issues, we innovatively propose Boosting Class-specific Face Privacy Protection with Stochastic Mask Regularized Adaptive Momentum (SMRAM) for privacy protection. This approach introduces Stochastic Mask Regularized Adaptive Momentum, which dynamically integrates fine-grained features and global consistency features through Adaptive Momentum (AM). Additionally, Stochastic Mask Regularization (SMR) increases the diversity of the perturbation distribution, thereby capturing richer generic features. When fine-grained features dominate, AM refines the generation direction by incorporating generic features and stabilizes the generation process through momentum accumulation, thereby improving the success rate of privacy protection. Furthermore, SMR forces the model to learn more generic features, enhancing the generalization of the perturbations. Extensive experiments conducted on two benchmark datasets demonstrate that the proposed SMRAM outperforms the state-of-the-art approaches, achieving a higher success rate in privacy protection for same-identity images under black-box recognition models.**

***Index Terms*—Privacy Protection, Adversarial Example, Class-Specific**

I. Introduction

DEEP learning has achieved significant progress in the field of scientific research [1], with applications spanning various domains such as entertainment, travel, and security. In particular, face recognition systems have made remarkable strides with the advancement of deep learning networks [2]. While these systems provide considerable convenience to society, their misuse presents a serious threat to public privacy. As individuals increasingly share personal photos on social media, which contain sensitive information, face recognition systems can be exploited to steal identities, uncover relationships, and facilitate fraud and other malicious activities.

To overcome the limitations of simple blurring or pixelation for facial images [3], researchers have focused on developing methods that protect personal privacy while preserving the usability of images and videos. Several studie have explored removing identity information from images prior to data sharing while retaining essential facial features to prevent accurate identification by facial recognition systems. However, de-identified images often display significant visual differences from the original [4] or introduce unnatural regions and unwanted visual artifacts [5], [6]. Other research [7], [8] has proposed generating subtle adversarial perturbations to protect the privacy of original facial images, resulting in adversarially crafted images designed to confuse recognition models. Current research identifies two main strategies for adversarial perturbations in privacy protection: one that customizes image-specific perturbations for each facial image [9], and another that uses a universal adversarial perturbation approach [10]. The image-specific method [11], [12] involves multiple optimization iterations to find the best perturbation, providing strong privacy protection by causing misclassification of adversarial examples across various models. However, this method requires significant computational resources. In contrast, universal privacy perturbations only require the generation of a single universal perturbation to protect all facial images. While this method saves computational resources, its ability to protect individual face images is somewhat limited due to the lack of fine-grained optimization for personal features.

Currently, Zhong et al. [13] proposed generating a universal privacy shield for all facial images of each individual, where only one shield is created per user and applied to all their images or videos, saving time and computational resources compared to image-dependent schemes. This method offers better privacy protection within the same class of images. Class-specific privacy protection generates perturbations by learning the facial features of each identity. However, due to factors like micro-expressions, images of the same identity often show high local similarity, causing perturbations to rely too much on specific regions [14], leading to overfitting and limited generalization. While increasing the number of images may introduce some variation, it doesn't significantly enhance sample diversity. Balancing the sensitivity of face recognition models to fine-grained features with the generalization of global consistency features remains a challenge.

To address the above problems, this paper proposes a novel approach called Boosting Class-specific Face Privacy Protection with Stochastic Mask Regularized Adaptive Mo-

mentum (SMRAM). SMRAM dynamically combines fine-grained features and globally consistent features through Adaptive Momentum (AM) and Stochastic Mask Regularization (SMR), generating class-specific adversarial perturbations. Specifically, to prevent the perturbation from overly relying on locally specific regions, SMR is introduced to diversify the adversarial perturbation by altering the attention distribution and capturing generic features. To further refine the perturbation generation process, AM allows dynamic adjustment between fine-grained features and generic features. By combining the attack strength of fine-grained features with the generalization ability of generic features and utilizing momentum accumulation to stabilize the perturbation optimization direction, SMRAM enhances the privacy protection ability of adversarial perturbations for specific classes. Our framework is shown in Figure 1, illustrating the overall structure and key components. In summary, the main contributions of this paper are:

- We propose a novel face privacy protection approach called SMRAM, which generates a universal adversarial perturbation applicable to all images of a given individual, thereby improving generalization.
- SMRAM dynamically balances fine-grained and generic features via AM strategy and utilizes SMR to diversify the perturbation distribution.
- Extensive experiments on black-box face recognition models demonstrate the merits of the proposed method. SMRAM can generate perturbations with higher privacy protection success rate under the same identity data than state-of-the-art approaches.

II. Related Work

A. Face privacy protection

The research on protection of face privacy has also been carried out for a long time. In order to protect the security of face data, some previous work proposed facial de-identification [**?**], [6] to remove factors that can expose identity information from face pictures. In order to replace the early attempts of fuzzy images, the k-anonymity protection model [15] transforms each face in a closed set into the average of the nearest K identities in the set. In recent years, with the continuous development of the generative network model, a large number of related works have also appeared in the field of face de-identification [5], [6]. Some recent studies use the generative network model to generate synthetic images that can replace the original images, and these synthetic images do not contain any identity information from the original face.

B. Adversarial Attacks

In earlier studies, it was found that although deep neural networks are quite effective in many fields today, they are also vulnerable to adversarial examples [9], [16]. Previous work can be simply divided into two types: one is the image-dependent adversarial perturbation, and the other is universal adversarial perturbation. Image-dependent adversarial perturbation [17], [18] operates on a single input image. It feeds the current image into the target model to obtain the gradient and iteratively optimizes the perturbation using this gradient information. The adversarial perturbation generated by separate optimization is only valid for the current input picture. Universal adversarial perturbation [19], [20] is an image-independent perturbation generation scheme capable of successfully attacking most recognition models across a substantial number of natural images. During generator training, Zhao et al. [20] incorporates minimized alternative model's maximum discrepancy to produce generalized adversarial instances. Zhong et al. [13] proposed class-specific universal adversarial perturbations, which customize a generic adversarial perturbation for each individual using a specific set of images to safeguard privacy.

III. Class-Specific Privacy Protection

A. Problem Formulation

The goal of privacy protection is to ensure the recognition model cannot accurately classify face images. Perturbations from untargeted adversarial attacks achieve this for human faces. The class-specific perturbation ΔX causes all images $(X^k = X_1^k, X_2^k, \ldots, X_i^k, \ldots)$ of the source identity k to be misclassified. Specifically, we aim to find ΔX such that

$$D(f(X_i^k + \Delta X), f_{X^k}) > t, \quad \|\Delta X\|_\infty < \varepsilon, \tag{1}$$

where $f(X_i^k) \in \mathbb{R}^d$ is the feature of image X_i^k from the model $f(\cdot)$, f_{X^k} is the feature subspace for identity k, and t is the distance threshold. $D(x_1, x_2)$ is the distance between points, usually Euclidean or cosine, and ε limits the perturbation's pixel amplitude.

In this paper, SMRAM utilizes both generic feature information from various images of the same identity and fine-grained features to improve adversarial perturbations. The Subspace Loss Calculation Module (SLC) computes the loss between the subspace and adversarial samples. The l_{NR} from non-regularized data captures fine-grained features, while l_{SR} from stochastic regularized data captures generic features. These losses effectively represent different feature types. Additionally, Adaptive Momentum (AM) is used to balance gradients and stabilize perturbation optimization, improving accuracy.

B. Subspace Loss Calculation Module

To generate identity-specific adversarial perturbations, we model the feature space f_{X^k} using the convex hull. For each face $f(X_i^k)$, we minimize the distance between the weighted linear combination of the deep feature matrix F^k from all faces of the same identity and the perturbed face features, obtaining the corresponding coefficient α_i^k. Finally, we obtain the coefficient vector A^k for all faces of the current identity during the training process. Once the coefficient matrix associated with the current source identity image is determined, the Subspace Loss Calculation module (SLC) can be used to measure the distance between the adversarial samples and the feature space, denoted as l_{NR}.

$$l_{NR} = \sum_{i=1}^{n_k} \|F^k A_i^k - f(X_i^k + \Delta X)\|, \tag{2}$$

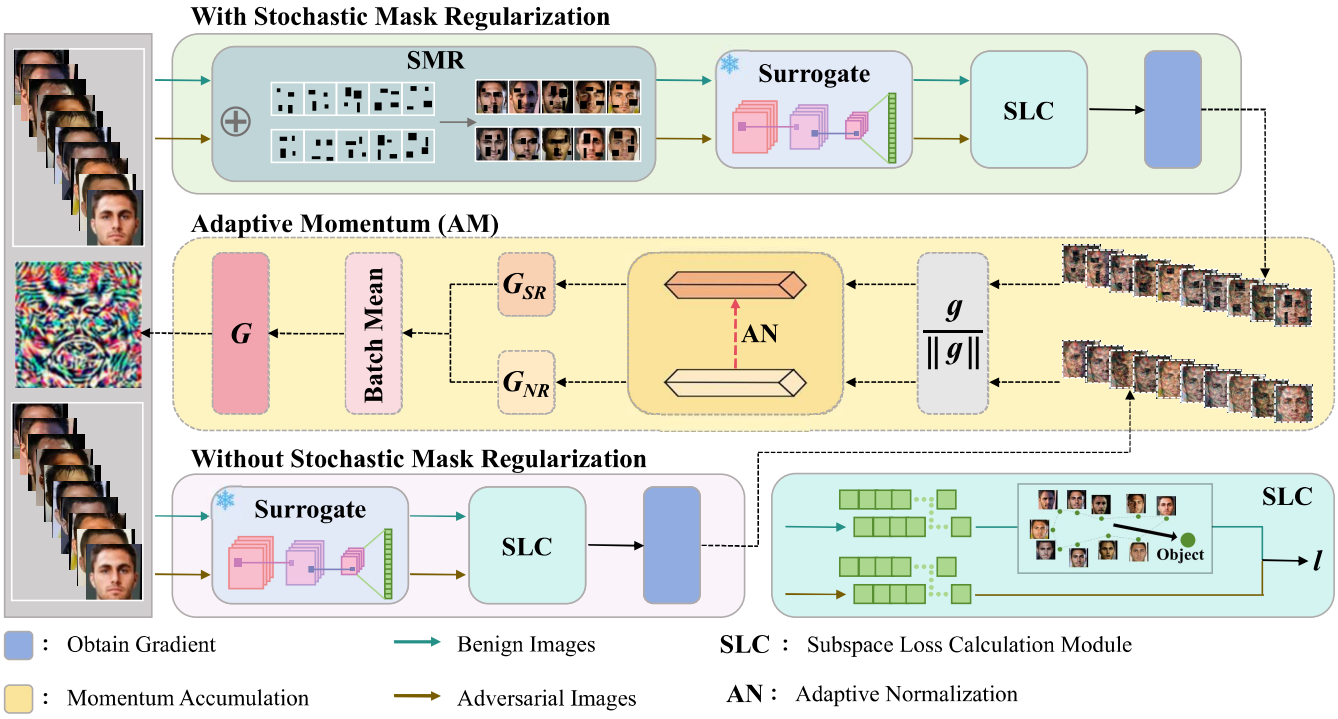

Fig. 1. SMRAM generates class-specific generic perturbations for each identity. The model employs the AM strategy to balance fine-grained features and generic features, ensuring the stability of the optimization process, and extracts generic feature information through SMR.

the gradient on the non-regularized data can be calculated from l_{NR}:

$$g_{NR} = \nabla_{X_i^k + \Delta X} l_{NR}, \tag{3}$$

C. Stochastic Mask Regularization

To enhance class-specific universal adversarial perturbations with limited facial data, we further introduce a Stochastic Mask Regularization (SMR) based on l_{SR}, leveraging the sample space. The method proposed in this paper incorporates SMR at corresponding locations in both benign samples and adversarial samples, thereby altering the attention distribution of the face recognition model toward the images. The perturbations are forced to learn more generic and universal features, which provides more information for optimizing adversarial perturbations. The processing of images by SMR is represented by $S(\cdot)$.

$$S(x) = (1 - \mathcal{A}) \odot x + \mathcal{A} \odot M \tag{4}$$

where M is an basic mask, and $\odot$ is the element-wise multiplication. $\mathcal{A}$ is a position matrix used to select the local locations in the image to be covered.

After recalculating the coefficient matrix $A^{k'}$ corresponding to the face images with SMR applied, the l_{SR} can be calculated from the processed original image and attack sample as follows:

$$l_{SR} = \sum_{i=1}^{n_k} \|S(F^k)A_i^{k'} - f(S(X_i^k + \Delta X))\| \tag{5}$$

the gradient of the overlaid image can be calculated as

$$g_{SR} = \nabla_{S(X_i^k + \Delta X)} l_{SR}, \tag{6}$$

D. Adaptive Momentum

When integrating fine-grained features and generic features from images of the same identity, we typically sum their gradients to form a composite gradient. In the traditional momentum optimization framework, momentum smooths the optimization trajectory and suppresses fluctuations by accumulating historical gradients, thereby improving stability and convergence. However, simple weighted fusion fails to capture the dynamic variations and nonlinear dependencies between gradients, limiting further improvements in overall performance.

To address this, we propose an adaptive momentum accumulation scheme. For the loss function l_{NR} corresponding to non-regularized data, we accumulate its gradients via a momentum mechanism to smooth fluctuations during updates, preserve historical gradient information, and effectively reduce random disturbances during iterations. Furthermore, for the loss function l_{SR} corresponding to regularized data, we introduce an adaptive normalization operation AN to balance the scale and directional differences between fine-grained feature gradients and global feature gradients, promoting coordination and complementarity between the two gradient types during updates. The formulation is expressed as follows:

$$G_{N+1}^{NR} = \mu \cdot G_N^{NR} + \frac{g_{NR}}{\|g_{NR}\|}, G_{N+1}^{SR} = AN(G_{N+1}^{NR}, g_{SR}), \tag{7}$$

$$G_{N+1} = \beta_1 \cdot G_{N+1}^{NR} + \beta_2 \cdot G_{N+1}^{SR}, \tag{8}$$

where G_N^{NR} and G_N^{SR} represent the accumulated momenta corresponding to the two loss functions at iteration N. The decay factor μ is conventionally fixed to 1, following the setting in [13]. The adjustable weighting parameters β_1 and β_2 perform a weighted fusion of the two types of gradient information into a composite gradient G_{N+1}, which is then used to update the adversarial perturbations.

And $AN(G, g)$ is an adaptive normalization operation that takes two gradients as input: one is the accumulated momentum gradient G, and the other is the gradient term g to be adjusted. The operation first calculates the range of the accumulated momentum gradient, the difference between its maximum and minimum values. It then linearly maps each gradient value from the second gradient term into this range, thereby aligning it with the previous gradient in terms of scale. The final output is the normalized gradient.

$$AN(G,\, g) = \min(G) + \frac{\max(G) - \min(G)}{\max(g) - \min(g)} \cdot (g - \min(g))\,, \tag{9}$$

where $G = G_{N+1}^{NR}$ represents the momentum-accumulated gradient from the non-regularized loss, which carries historical update information and tends to be stable in scale. And $g = g_{SR}$ is the raw gradient from the regularized loss, which may exhibit different magnitude characteristics due to the stochastic masking process. By mapping g_{SR} into the value range of G_{N+1}^{NR}, we explicitly align the two gradient streams in scale. The Adaptive Momentum mechanism, by introducing the $AN(G, g)$ operation, enables adaptive alignment in gradient direction, mitigates fluctuations during gradient updates, and enhances the stability of the optimization process in generating adversarial samples. Additionally, the $AN(G, g)$ operation improves the reproducibility of the optimization process, ensuring stable and consistent model performance under identical perturbation conditions.

SMRAM leverages a substantial amount of generic feature information from multiple images of the same identity, in conjunction with fine-grained features, to enhance the effectiveness of generalized adversarial perturbations. The loss between the feature subspace and adversarial samples is computed via the Subspace Loss Calculation Module (SLC). The l_{NR} loss, derived from non-regularized data, captures rich fine-grained feature information, while the l_{SR} loss, from stochastic regularized data, emphasizes broader, generic features. By utilizing these two distinct losses, we effectively model their respective feature representations. Furthermore, AM is employed to balance the gradients from both losses and stabilize the optimization of perturbations, improving both accuracy and convergence. This dual-loss approach ensures that fine-grained and generic features are appropriately integrated, promoting both precision and generalization in adversarial perturbation generation.

TABLE I

COMPARISON OF SMRAM AND CLASS-SPECIFIC PRIVACY PERTURBATION METHODS COMBINED WITH MOMENTUM ON THE SOURCE MODEL USING THE PRIVACY-COMMONS AND PRIVACY-CELEBRITIES DATASETS, WITH RESULTS REPORTED AS ATTACK SUCCESS RATES (%).

Dataset	Method	Target			
		ArcFace	CosFace	SFace	MobileNet
Privacy-Commons	GAP	30.7	19.7	23.9	33.6
	FI-UAP	80.2	72.0	77.9	82.7
	FI-UAP+	80.0	73.3	78.7	83.0
	OPOM	81.2	73.6	79.1	83.6
	SMR-M	82.3	75.5	80.4	84.6
	SMRAM	**83.3**	**76.5**	**81.9**	**85.3**
Privacy-Celebrities	GAP	41.5	31.7	35.0	53.2
	FI-UAP	62.7	52.9	60.3	67.8
	FI-UAP+	64.2	54.6	61.9	69.3
	OPOM	65.7	55.5	63.0	70.1
	SMR-M	66.7	56.8	64.4	70.9
	SMRAM	**68.5**	**58.8**	**65.9**	**72.2**

TABLE II

COMPARISON OF CLASS-SPECIFIC PRIVACY PERTURBATION METHODS ON THE SOURCE MODEL USING THE PRIVACY-COMMONS AND PRIVACY-CELEBRITIES DATASETS, WITHOUT MOMENTUM, WITH RESULTS REPORTED AS ATTACK SUCCESS RATES (%).

Dataset	Method	Target			
		ArcFace	CosFace	SFace	MobileNet
Privacy-Commons	FI-UAP	72.3	63.5	70.3	73.9
	FI-UAP+	76.9	69.2	75.1	78.3
	OPOM	78.0	70.2	76.1	79.2
	SMRAM-SMR	**79.6**	**72.0**	**77.8**	**80.7**
Privacy-Celebrities	FI-UAP	56.5	45.2	52.9	60.2
	FI-UAP+	62.4	51.8	59.8	66.2
	OPOM	63.8	53.6	61.3	67.0
	SMRAM-SMR	**65.0**	**54.5**	**62.6**	**68.5**

IV. EXPERIMENTS

A. Comparison Methods

To evaluate the effectiveness of our approach, we compare it with representative methods for class-specific universal adversarial perturbations in face recognition. FI-UAP generates universal perturbations by applying DeepFool [21] to multiple images of the same identity and aggregating minimal perturbations, while FI-UAP+ enhances this with in-class aggregation to better capture intra-class consistency. GAP [22] employs a generator trained with feature-level loss to produce transferable adversarial perturbations. OPOM [13] learns identity subspaces and iteratively optimizes perturbations, yielding class-specific universal adversarial perturbations.

B. Experimental Settings

In this paper, we employ attack success rates as the evaluation metric for privacy protection performance. For each identity, M images are prepared, with interference images incorporated to construct the gallery, while the remaining $M-1$ are used as probes. During testing, recognition accuracy is measured by ranking feature distances between gallery

and probe images, and the Top-1 protection success rates is reported.

To evaluate the effectiveness of adversarial perturbations, we employed the Privacy-Commons [23] and Privacy-Celebrities [24] datasets. Each dataset contains 500 individuals with approximately 5,000 facial images for training. The diversity and scale of these datasets help validate the method's generalization ability across different types of face recognition tasks, ensuring that the proposed approach maintains a high level of privacy protection in various datasets and real-world application scenarios. In the experiments, a modified ResNet-50 [25], [26], trained on the CASIA-WebFace dataset [27], was adopted as the surrogate model. For the target face recognition models, CosFace [28], ArcFace [26], and SFace [29] were implemented with different loss functions, while MobileNet [30] served as an alternative network architecture for comparison.

C. Effectiveness of SMRAM

We conducted a comprehensive evaluation of SMRAM, which integrates AM and SMR, on the Privacy-Commons and Privacy-Celebrities datasets and performed an in-depth comparison with different methods. To ensure a fair comparison, momentum was also applied to the competing methods FI-UAP, FI-UAP+, and OPOM. To assess the contribution of AM, we further tested SMR separately under the momentum setting, where L_{NR} and L_{SR} were weighted by (β_1, β_2) and updated using a unified momentum scheme, resulting in SMR-M.

As shown in Table I, under the momentum setting, SMR-M achieves a higher average attack success rate than OPOM on both datasets. This confirms that SMR enhances the privacy protection of adversarial perturbations by enriching the perturbation distribution, which provides generic information and, together with fine-grained details, improves the overall effectiveness. Moreover, based on SMR, SMRAM with integrated AM exceeds SMR-M and OPOM by **1.3%** and **2.6%**, respectively, on both datasets. The performance gains, though numerically modest, are statistically significant, as SMRAM consistently outperforms all baselines across every metric and target model on both datasets. This systematic superiority validates the efficacy of the AM mechanism and highlights the synergistic effect of integrating AM with SMR in our framework. However, the integration of AM and SMR may slightly increase computational complexity. The additional overhead from dynamically adjusting gradients and optimizing perturbations is relatively small and does not significantly affect efficiency.

In Figure 3, attention distributions are compared before and after applying SMR. Before regularization, attention is highly localized, indicating a reliance on local features. After regularization, attention becomes more dispersed and enriched, showing that the model captures more comprehensive and diverse features. This confirms that SMR enhances attention diversity and global information use through structured constraints, underscoring regularization's role in improving model generalization.

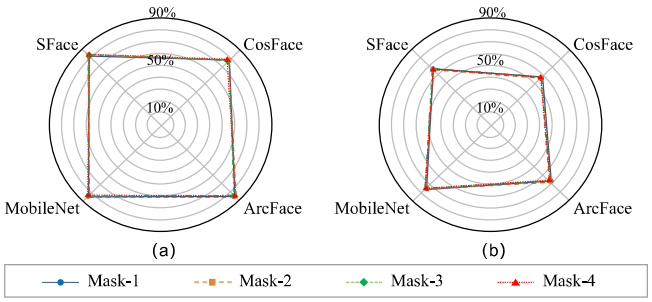

Fig. 2. The attack success rates of adversarial perturbations across different numbers of masks (1 to 4) on two datasets: (a) Privacy-Commons and (b) Privacy-Celebrities.

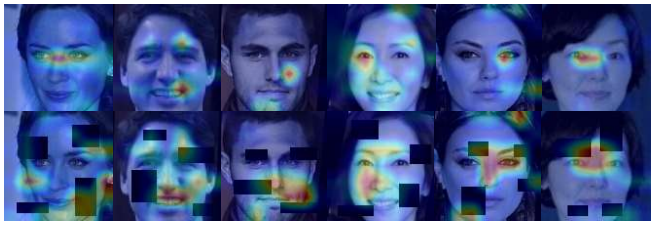

Fig. 3. The figure compares the attention maps before and after SMR regularization: the first row (before regularization) shows a higher concentration of attention, while the second row (after regularization) exhibits a significantly richer distribution. This validates that SMR guides the model to capture more comprehensive feature information through its constraint mechanism.

D. Ablation Study

To evaluate SMR and AM, we conducted ablation experiments. In the proposed method, SMR is used to provide generic feature information. To assess its effectiveness, we compared its performance with FI-UAP, FI-UAP+, and OPOM. To ensure a fair comparison, none of the methods employed transferability-enhancing strategies like momentum, and SMR did not use either standard or AM. During training, SMR optimizes the perturbation by combining regularized and non-regularized losses. The experimental results in Table II show that SMR improves the average attack success rate by **2.2%** over OPOM on two datasets. This statistically significant improvement demonstrates that SMR can more effectively leverage generic information, not only markedly enhancing identity-specific privacy protection performance but also improving the generalization capability of adversarial samples. The results validate the effectiveness of the proposed regularization mechanism in class-sensitive tasks from both dimensions of effect size and consistency.

Additionally, we further studied the effect of stochastic mask count on perturbation performance. Results confirm that our method maintains reliable privacy protection across varying mask numbers. Too few masks may insufficiently disrupt feature extraction, while too many can lead to excessive information loss. Our stochastic masking mechanism, which introduces randomness in mask size and location, ensures perturbation reliability across different mask quantities. Experiments with varying numbers of masks (1, 2, 3, and 4) in

Figure 2 show that the number of masks has a minimal impact on privacy protection success. To ensure fairness, we used four masks in all comparative experiments, reinforcing the stability and generalizability of our approach.

V. Conclusion

In this paper, we propose a novel method for face privacy protection, termed Boosting Class-specific Face Privacy Protection with Stochastic Mask Regularized Adaptive Momentum (SMRAM). This method uses Adaptive Momentum (AM) to dynamically balance generic features and fine-grained features, while utilizing momentum accumulation to stabilize the perturbation optimization direction, thus avoiding local optima and enhancing the success rate of privacy protection. To further facilitate the learning of generic features and alleviate overfitting to narrow regions, SMRAM employs Stochastic Mask Regularization (SMR) to diversify the perturbation distribution, capturing more generic feature information and thereby improving its generalization ability. Extensive experiments on two benchmark datasets demonstrate that SMRAM outperforms existing methods, generating more generalizable results for the same identity data. Ablation studies validate the universality and stability of the proposed method. Future work will focus on integrating the latest generative models to further enhance the performance of the algorithm.

VI. Acknowledgment

This work was supported in part by the National Key R&D Program of China (2023YFE0116300), in part by the National Natural Science Foundation of China (Grant No.62006097, U1836218).

References

[1] A. Zhao, D. Huang, Q. Xu, M. Line, Y.-J. Liu, and G. Huang, "Expel: Llm agents are experiential learners," pp. 19 632–19 642, 2024.

[2] M. K. Tellamekala, S. Amiriparian, B. W. Schuller, E. André, T. Giesbrecht, and M. Valstar, "Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 2, pp. 805–822, 2024.

[3] M. Boyle, C. Edwards, and S. Greenberg, "The effects of filtered video on awareness and privacy," in *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, ser. CSCW '00. New York, NY, USA: Association for Computing Machinery, 2000, p. 1–10.

[4] J. Liu, Z. Zhao, P. Li, G. Min, and H. Li, "Enhanced embedded autoencoders: An attribute-preserving face de-identification framework," *IEEE Internet of Things Journal*, vol. 10, no. 11, pp. 9438–9452, 2023.

[5] J. Lin, Y. Li, and G. Yang, "Fpgan: Face de-identification method with generative adversarial networks for social robots," *Neural Networks*, vol. 133, pp. 132–147, 2021.

[6] G. Feng, T. Yan, and J. Yang, "Attribute-consistency reversible pedestrian de-identification in intelligent transportation," *IEEE Transactions on Vehicular Technology*, vol. 74, no. 2, pp. 2187–2197, 2025.

[7] Y. Li, C. Hu, R. Wang, and X. Wu, "Hierarchical feature transformation attack: Generate transferable adversarial examples for face recognition," *Applied Soft Computing*, vol. 172, p. 112823, 2025.

[8] C. Hu, Z. He, and X. Wu, "Query-efficient black-box ensemble attack via dynamic surrogate weighting," *Pattern Recognition*, vol. 161, p. 111263, 2025.

[9] Y. Yang, Y. Huang, M. Shi, K. Chen, and W. Zhang, "Invertible mask network for face privacy preservation," *Information Sciences*, vol. 629, pp. 566–579, 2023.

[10] H. Gao, H. Zhang, X. Zhang, W. Li, J. Wang, and F. Gao, "Enhanced covertness class discriminative universal adversarial perturbations," *Neural Networks*, vol. 165, pp. 516–526, 2023.

[11] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: Protecting privacy against unauthorized deep learning models," in *29th USENIX security symposium (USENIX Security 20)*, 2020, pp. 1589–1604.

[12] J. Zhang, J. Sang, X. Zhao, X. Huang, Y. Sun, and Y. Hu, "Adversarial privacy-preserving filter," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1423–1431.

[13] Y. Zhong and W. Deng, "Opom: Customized invisible cloak towards face privacy protection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3590–3603, 2022.

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[15] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *Proceedings of the 31st International Conference on Very Large Data Bases*, ser. VLDB '05. VLDB Endowment, 2005, p. 901–909.

[16] M. Duan, Y. Qin, J. Deng, K. Li, and B. Xiao, "Dual attention adversarial attacks with limited perturbations," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 10, pp. 13 990–14 004, 2024.

[17] Z. Su, D. Zhou, N. Wang, D. Liu, Z. Wang, and X. Gao, "Hiding visual information via obfuscating adversarial perturbations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4356–4366.

[18] K. Liang and B. Xiao, "Styless: Boosting the transferability of adversarial examples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 8163–8172.

[19] J. Weng, Z. Luo, Z. Zhong, D. Lin, and S. Li, "Exploring non-target knowledge for improving ensemble universal adversarial attacks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 2768–2775.

[20] A. Zhao, T. Chu, Y. Liu, W. Li, J. Li, and L. Duan, "Minimizing maximum model discrepancy for transferable black-box targeted attacks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 8153–8162.

[21] Y. Shi, Y. Han, Q. Hu, Y. Yang, and Q. Tian, "Query-efficient black-box adversarial attack with customized iteration and sampling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2226–2245, 2022.

[22] O. Poursaeed, I. Katsman, B. Gao, and S. Belongie, "Generative adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4422–4431.

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

[24] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 87–102.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

[26] 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*, 2019, pp. 4690–4699.

[27] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[28] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.

[29] Y. Zhong, W. Deng, J. Hu, D. Zhao, X. Li, and D. Wen, "Sface: Sigmoid-constrained hypersphere loss for robust face recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 2587–2598, 2021.

[30] A. G. Howard, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.