CSVAR: Enhancing Visual Privacy in Federated Learning via Adaptive Shuffling Against Overfitting

Zhuo Chen, Zhenya Ma, Yan Zhang, Donghua Cai, Ye Zhang, Qiushi Li, Yongheng Deng, Ye Guo, Ju Ren and Xuemin (Sherman) Shen

***Abstract*—Although federated learning preserves training data within local privacy domains, the aggregated model parameters may still reveal private characteristics. This vulnerability stems from clients' limited training data, which predisposes models to overfitting. Such overfitting enables models to memorize distinctive patterns from training samples, thereby amplifying the success probability of privacy attacks like membership inference. To enhance visual privacy protection in FL, we present CSVAR(Channel-Wise Spatial Image Shuffling with Variance-Guided Adaptive Region Partitioning), a novel image shuffling framework to generate obfuscated images for secure data transmission and each training epoch, addressing both overfitting-induced privacy leaks and raw image transmission risks. CSVAR adopts *region-variance* as the metric to measure visual privacy sensitivity across image regions. Guided by this, CSVAR adaptively partitions each region into multiple blocks, applying fine-grained partitioning to privacy-sensitive regions with high region-variances for enhancing visual privacy protection and coarse-grained partitioning to privacy-insensitive regions for balancing model utility. In each region, CSVAR then shuffles between blocks in both the spatial domains and chromatic channels to hide visual spatial features and disrupt color distribution. Experimental evaluations conducted on diverse real-world datasets demonstrate that CSVAR is capable of generating visually obfuscated images that exhibit high perceptual ambiguity to human eyes, simultaneously mitigating the effectiveness of adversarial data reconstruction attacks and achieving a good trade-off between visual privacy protection and model utility.**

***Index Terms*—Anti-Overfitting, Visual privacy protection, Image shuffling, Federated learning**

I. Introduction

Federated learning (FL [1]) has emerged as a promising decentralized learning paradigm that enables multiple participants to collaboratively train a shared model without sharing their raw data. Instead of centralizing sensitive data in the cloud, FL allows clients to train locally and only exchange model updates (e.g., gradients or weights) with a central server for aggregation. This framework enhances privacy by preserving training data within local privacy domains, while maintaining model performance. Thus, FL is particularly valuable for vision-based applications like medical imaging analysis, facial recognition systems, and mobile photography services where protecting visual privacy and preventing reconstruction of identifiable images or extraction of sensitive visual features are paramount.

Despite its privacy-aware design, FL still faces two critical vulnerabilities for visual data protection. First, although clients only share model updates rather than raw images with the remote server, the typically small and non-IID local datasets [2] often lead to severe overfitting during local training. This causes model weights to encode excessive information about specific training samples, enabling adversaries to visually reconstruct private images through Model Inversion Attacks [3] (e.g., recovering patient faces from medical image models) or determine whether a given image belongs to training sets through Membership Inference Attacks [4]. Second, in vision-centric deployments (e.g., healthcare systems with CT/MRI scanners, smart cameras), a security gap exists between image collection sensors and computing nodes. The transmission of private raw images within local networks creates attack surfaces where images could be intercepted, undermining FL's end-to-end privacy guarantees for private data [5].

Differential privacy (DP) [6] has been widely adopted in FL to mitigate such privacy risks through introducing carefully calibrated random noise into training processes. For visual data protection, this randomness ensures each training iteration operates on effectively varied versions of the input images, alleviating the overfitting phenomenon. However, DP's noise-based protection operates in the high-frequency domain - while mathematically sound for membership privacy, the human eyes can easily filter such perturbations, leaving sensitive image features exposed [7]. What's worse, achieving strong visual privacy through DP often requires significant noise injection, which often degrades model utility unacceptably.

Building upon these limitations of DP, we explore the image shuffling mechanism that can generate obfuscated images for data transmission and each training epoch, addressing both overfitting-induced privacy leaks and raw image transmission risks. However, designing an effective shuffling strategy faces several nuanced challenges. First, there exists the non-uniform nature of visual privacy across different image regions: sensitive regions like faces or medical identifiers require stronger obfuscation than structural features like backgrounds. This

Both authors (Zhuo Chen and Zhenya Ma) contributed equally to this work.

Zhuo Chen and Ye Guo are with China Mobile Communications Group Co.,Ltd. (email: chenzhuoit@chinamobile.com, guoye@chinamobile.com)

Zhenya Ma, Yan Zhang, Donghua Cai, Qiushi Li (corresponding author) and Yongheng Deng are with the Department of Computer Science and Technology, Tsinghua University, China. (email: mzy23@mails.tsinghua.edu.cn, yan-zhan23@mails.tsinghua.edu.cn, caidonghua2003@gmail.com, lqs@tsinghua.edu.cn, dyh2024@tsinghua.edu.cn)

Ye Zhang is with the School of Computer Science, Beijing Information Science and Technology University, China. (email: yezhang@bistu.edu.cn)

Ju Ren is with the Department of Computer Science and Technology, Tsinghua University, China, and Zhongguancun Laboratory, Beijing, China. (email: renju@tsinghua.edu.cn)

Xuemin (Sherman) Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. (email: sshen@uwaterloo.ca)

necessitates an adaptive shuffling granularity of a given region: overly fine-grained shuffling, like pixel-level shuffling, maximally alleviates overfitting but destroys essential spatial structures needed for model learning, whereas overly coarse shuffling risks insufficient privacy protection. We need to divide the region into smaller blocks, shuffle between them, and preserve the pixels within the block. A one-granularity-fits-all shuffling approach would either over-protect unimportant areas or under-protect sensitive ones, motivating the need for spatially adaptive transformations to make the trade-off. Second, we need a principled metric to measure a region's visual privacy sensitivity (human-recognizable features) and model overfitting risk. Without such guidance, naive shuffling implementations either compromise privacy through insufficient obfuscation or degrade model utility unacceptably via excessive shuffling. Third, while conventional pixel-wise spatial shuffling provides basic spatial obfuscation, it fails to disrupt color distributions - a critical vulnerability as attackers can reconstruct sensitive features through chromatic analysis.

To address the challenges outlined above, we propose CSVAR (Channel-Wise Spatial Image Shuffling with Variance-Guided Adaptive Region Partitioning), a novel anti-overfitting mechanism that employs dynamic region-adaptive shuffling. By preventing model overfitting through variance-guided image shuffling, CSVAR simultaneously prevents overfitting-induced visual privacy leakage and client-side data transmission under FL.CSVAR will first divide the images into regions and calculate each region's variance. Guided by variances, CSVAR will partition the region into smaller blocks adaptively and shuffle between blocks both in the spatial domain to destroy spatial visual features and the chromatic channel space to disrupt color distributions. Specifically, CSVAR will partition a privacy-sensitive region with high variance into small blocks to enhance privacy protection. Meanwhile, for a privacy-insensitive region with low variance, CSVAR will partition it with much bigger blocks to preserve model utility. CSVAR ensures that: 1) private training images are protected during client-side transmission by transmitting obfuscated versions instead of raw images, 2) different shuffled versions are used in different epochs to alleviate overfitting and overfitting-induced privacy leakage, and 3) the adaptive shuffling approach maintains a careful balance between visual privacy protection and model utility throughout the federated learning process.

Our contributions are summarized as follows:

- We propose CSVAR, a novel image shuffling framework to prevent visual privacy leakage from overfitting and client-side data transmission under Federated Learning.
- We adopt region-variance as the metric to quantify the visual privacy sensitivity of different regions of an image and guide the shuffling granularity.
- We adaptively partition each image region into smaller blocks with different granularity guided by region-variance, and then shuffle between blocks in both the spatial domains and chromatic channels.
- We conduct extensive experiments on real-world datasets, demonstrating that CSVAR can generate visually obfuscated images that exhibit high perceptual ambiguity to human eyes, mitigate the effectiveness of adversarial data reconstruction attacks, and achieve a good trade-off between visual privacy protection and model utility.

II. Threat Model and Design Goals

This section formalizes our system model and threat model, followed by our design goals.

A. System Model

We consider a conventional federated learning framework comprising a central server and multiple distributed clients, composed of two functionally distinct components: local data collection and computation nodes. This architecture mirrors a lot of real-world deployments, particularly in privacy-sensitive domains like healthcare, where medical imaging scanners (CT/MRI, lacking computational capabilities) must send collected data to local computation nodes through internal networks to perform local model training.

During each training epoch, the system operates through several key phases. Initially, data collection nodes gather raw local training data through their sensing capabilities, then transfer private data to co-located computation nodes via internal networks. The computation nodes subsequently perform local model training using their private datasets before submitting parameter updates to the central server. Following aggregation of clients' updates, the server distributes the updated global model back to all participating clients for the next training epoch.

B. Security Model

We focus on two key attack scenarios for image data protection in the above system model. First, attackers can intercept client-side data transmissions between data collection nodes and computation nodes during internal network transfers. Second, attackers can access the server's global model weights. Here, the attacker can be the curious-but-honest servers that honestly perform federated learning tasks but attempt to extract private data from model weights by employing techniques such as GAN-based data reconstruction attacks to reconstruct private training images or membership inference attacks to identify the ownership of the data used in the training process. These scenarios cover both raw data exposure during local transfers on the client side and potential privacy leaks through model weights on the server side.

C. Design Goals

Our framework aims to achieve three fundamental objectives that alleviate overfitting, and preserve the visual privacy of image data while retaining the utility of the model:

Secure data transmission between data collection and computation nodes on the client side. Our framework should provide robust obfuscation to the private image data to avoid privacy leakage even if the attacker can access obfuscated data.

Defend against visual privacy leakage from the overfit model on the server side. Our proposed method should

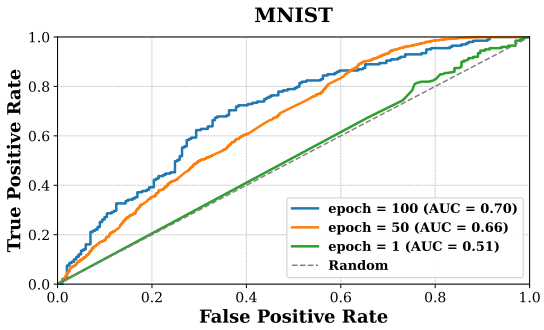

Fig. 1. ROC results of Membership Inference Attacks Varying Degrees of Overfitting. Low AUC means low attack Success Rate.

address the overfitting-induced privacy leakage on the private client data: (1)preventing data reconstruction attacks from reconstructing private images and (2) resisting membership inference attacks that attempt to identify training data participation.

Achieve satisfactory trade-off between visual privacy protection and model utility. Our framework should aggressively protect privacy-sensitive areas while maintaining structural integrity in privacy-insensitive regions, to minimize accuracy degradation while meeting privacy requirements.

III. Motivation

A. Overfitting-Induced Privacy Leakage in FL

Federated learning systems are particularly vulnerable to overfitting due to the typically small and non-IID nature of local client datasets. As training progresses, this overfitting causes model weights to encode increasingly sensitive features about private training samples, and we verify the growth of privacy leakage through **membership inference attacks(MIA) [4]**. Taking advantage of the model weights after training, MIA aims to determine whether a given data point belongs to the training dataset. Figure 1 demonstrates that moderately trained models have low MIA success rates (showing accuracy with low AUC near 0.5 similar to Random Guess), while heavily overfit models show high success rates (high AUC of 0.70).

This motivates our key insight: by shuffling training images for each epoch, we can ensure models encounter varying versions of the data, thereby simultaneously mitigating overfitting and reducing associated privacy leakage risks.

B. Region-Variance as Visual Privacy Indicator

Visual privacy protection requires recognizing that privacy exhibits an inherent spatial non-uniform nature: while privacy-sensitive regions(e.g., facial features) demand strong obfuscation, homogeneous backgrounds can tolerate lighter protection. An intuition that the variance of a given region can be used to measure privacy-sensitivity, based on the observation that sensitive areas typically exhibit higher pixel-value variance due to complex textures and edges, whereas uniform backgrounds show minimal variance.

To verify this intuition, our variance computation follows three steps: (1) partitioning the image into 14×14 regions

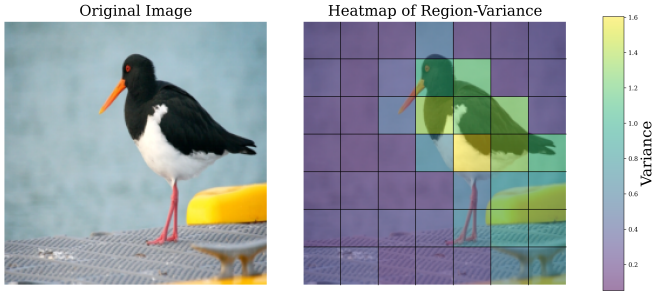

Fig. 2. Region-Variance of different regions in the image of a bird. Lighter region means a higher region-variance.

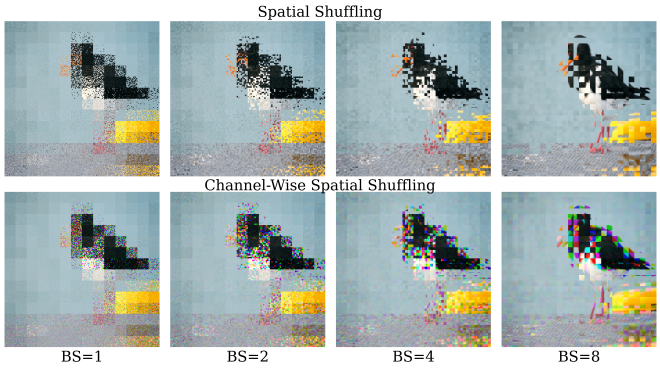

Fig. 3. Visual privacy protection effect under different block size(BS) with spatial-only shuffling and channel-wise spatial shuffling. Results show we can recognize fewer visual features when using a small BS with channel-wise spatial shuffling than a large BS with spatial-only shuffling.

(16×16 pixels each), (2) calculating per-channel variance across all pixels within each region, and (3) averaging variances from each channel. As shown in Figure 2, results show that high-variance regions (lighter areas in the heatmap) consistently correspond to semantically sensitive features (like the bird's beak/body in Figure 2). The strong alignment between variance and visual sensitivity confirms region-variance's suitability for guiding protection strategies, thus, we propose region-variance as a quantifiable metric for privacy sensitivity.

C. Shuffling Granularity Trade-off

The effectiveness of visual privacy protection through shuffling depends critically on shuffling granularity. As shown in Figure 3, we systematically explore this by the following steps: (1) dividing each 224×224 image into 14×14 base regions(16×16 pixels in each region), (2) further splitting each region into BS×BS blocks ($BlockSize(BS) \in [1, 2, 4, 8]$), and (3) shuffling between blocks while preserving intra-block pixels. Columns varying with different BS in figure 3 demonstrate that we can recognize more visual features with a large shuffle granularity(BS=8), which means we can get more visual privacy protection with small BS (e.g., BS=1). This is because pixels within each block will not be shuffled. Small BS provides stronger visual obfuscation by thoroughly disrupting spatial relationships, but at the cost of damaging semantically important structures needed for model learning. Conversely, large blocks maintain better model utility but leak more recognizable features. One key insight we obtain from this is that an effective shuffling strategy requires adaptive BS

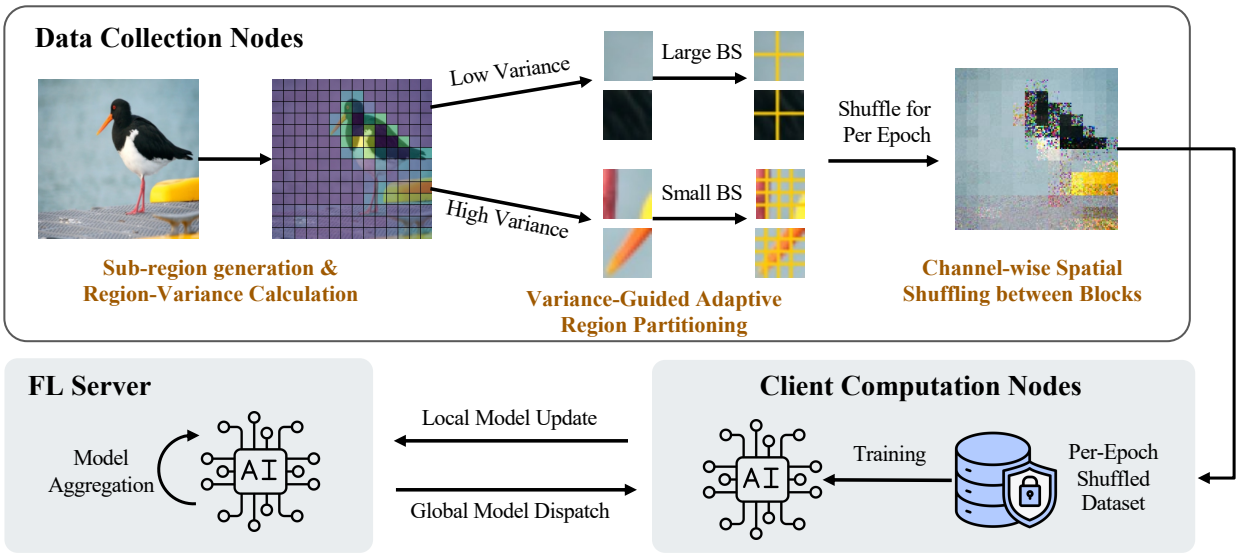

Fig. 4. The System Overview of CSVAR. Note that light color in the heatmap denotes a high variance region. BS means block size.

to partition different regions - using small BS for regions with high region-variances and large BS elsewhere.

Furthermore, the second row of Figure 3 shows the shuffled image after applying spatial shuffling to each RGB channel. Results show this channel-wise spatial shuffling provides additional visual privacy protection by disrupting color distributions, thereby preventing potential reconstruction attacks through chromatic analysis.

IV. System Design

A. System Overview

Building upon the motivation and design goals above, we present the system architecture of CSVAR. Figure 4 illustrates the end-to-end workflow of CSVAR. The process begins at the client data collection nodes, where raw private images are applied offline shuffling to generate distinct obfuscated versions for each training epoch and data transmission. First, CSVAR partitions each image into multiple regions and computes their region-variances as the privacy sensitivity metric. Guided by region-variance, CSVAR then adaptively partitions each region: privacy-sensitive regions with high variance will be divided into smaller blocks to maximize obfuscation, while low-variance regions with less sensitive features are split into larger blocks to better preserve model utility. Finally, CSVAR performs block-wise spatial shuffling, which shuffles between blocks both in the spatial domain to destroy spatial visual features and the chromatic channel space to disrupt color distributions. These shuffled images are then transmitted to the client computation nodes for local model training.

During the online federated learning phase, each training epoch follows a distributed workflow where participating computation nodes execute local model training on epoch-specific shuffled images. These clients subsequently transmit their local model updates to the central server. Upon receiving updates from all clients, the server employs a secure aggregation protocol to compute the global model update, which is then redistributed to participating nodes for the subsequent epoch. Notably, each training epoch incorporates distinct versions of shuffled image data to alleviate model overfitting.

This design ensures that: (1) Private images are protected during client-side transmission by sending obfuscated versions instead of raw images, (2) different shuffled versions are used in different epochs to alleviate overfitting and overfitting-induced privacy leakage, and (3) the adaptive approach maintains a careful balance between visual privacy protection and model utility throughout the federated learning process.

B. Design Details

1) Sub-region Generation & Region-Variance Calculation. CSVAR begins by partitioning each input image $I \in \mathbb{R}^{H\times W\times C}$ into non-overlapping regions $R_{i,j}$ of size $S \times S$ pixels, where S is determined adaptively based on image Height(H) and Width(W):

$$S = 2^{\lceil \log_2(\sqrt{\max(H,W)}) \rceil} \tag{1}$$

This choice of region size ensures: (1) sufficient granularity for privacy protection while maintaining recognizable local features ($S = 16$ for standard 224×224 images), (2) power-of-two region sizes enable natural binary partitioning(like $16 \to 8 \to 4$) into smaller blocks in the subsequent step.

For each region $R_{i,j}$, we compute its privacy sensitivity metric - the region-variance $RV^2_{i,j}$ - through:

$$RV^2_{i,j} = \frac{1}{C}\sum_{c\in C}\left[\frac{1}{|R_{i,j}|}\sum_{(x,y)\in R_{i,j}}(I_{x,y,c} - \mu_{i,j,c})^2\right] \tag{2}$$

where:

$I_{x,y,c}$: Pixel value at position (x, y) in channel c

$\mu_{i,j,c} = \frac{1}{|R_{i,j}|}\sum_{(x,y)\in R_{i,j}} I_{x,y,c}$ (Mean value in channel c)

$|R_{i,j}| = S^2$ (Number of pixels per region)

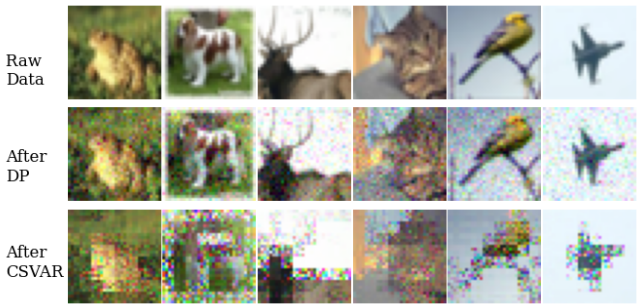

Fig. 5. Comparison of Visual obfuscation effect with different Protections.

2) Variance-Guided Adaptive Region Partitioning. Building upon the region-variance calculations, CSVAR implements an adaptive partitioning strategy that automatically adjusts protection strength based on each region's privacy sensitivity. Specifically, CSVAR will partition a privacy-sensitive region with high variance into small blocks to enhance privacy protection. Meanwhile, for a privacy-insensitive region with low variance, CSVAR will partition it with much bigger block sizes to preserve model utility.

CSVAR first computes the median variance $\tilde{RV}^2$ across all regions to establish the sensitivity threshold. For each region $R_{i,j}$:

$$\text{BlockSize} = \begin{cases} \lfloor S/4 \rfloor & \text{if } RV_{i,j}^2 > \tilde{RV}^2 \text{ (privacy-sensitive)} \\ \lfloor S/2 \rfloor & \text{if } RV_{i,j}^2 \leq \tilde{RV}^2 \text{ (privacy-insensitive)} \end{cases} \tag{3}$$

3) Channel-wise Spatial Shuffling between Blocks. After adaptive block partitioning, CSVAR performs two complementary shuffling operations to protect visual privacy. First, *spatial shuffling* randomly permutes the positions of all blocks within each region using different random seeds for each training epoch. This breaks spatial correlations between blocks while preserving pixel relationships within each block.

Second, *channel-wise shuffling* processes all channels independently: each block is decomposed into per-channel sub-blocks (e.g., R/G/B for color images), which are then shuffled across the region. For example, a block's first channel may relocate to the region's top-left while other channels scatter to different positions. This dispersion breaks channel correlations to disrupt color distributions, as adjacent positions now contain uncorrelated channels from different blocks.

V. EVALUATION

In this section, we conduct extensive experiments across various real-world datasets and models to evaluate CSVAR's effectiveness in enhancing visual privacy and preservation in model utility. We begin with the experimental setups, then present evaluation results to answer the following questions:

- Can CSVAR enhance visual privacy to generate visually obfuscated images that exhibit high perceptual ambiguity to human eyes and mitigate the effectiveness of adversarial data reconstruction attacks?
- Can CSVAR achieve a good trade-off between visual privacy protection and model utility?

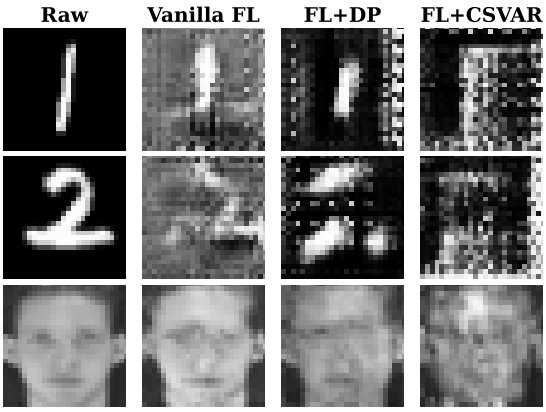

Fig. 6. Effect of GAN-based Data Reconstruction attack.

A. Experiment Setup

Models. We evaluate CSVAR using three representative CNN architectures selected to span the spectrum of modern computer vision applications. ResNet-50 serves as the base residual network, representing standard medium-scale models widely used in FL systems. MobileNet is included as the lightweight architecture optimized for edge devices with limited computational resources. ShuffleNet provides an additional efficiency-focused design point. This selection covers a range of model capacities (4.2M to 25.5M parameters) and computational requirements (0.6B to 4.1B FLOPs), ensuring thorough evaluation of CSVAR's compatibility across different neural network designs commonly deployed in computer vision applications under federated learning scenarios.

Datasets. Our experiments employ three benchmark datasets widely used in computer vision. CelebA Face Dataset (400 grayscale images of 40 subjects) evaluates the effectiveness of facial privacy protection. MNIST (70,000 handwritten digits) can assess preservation of basic structural features while preventing digit recognition. CIFAR-10 (60,000 color images across 10 categories) can test performance on more complex natural images with varied textures and compositions.

Baselines. We compare CSVAR against two baselines: (1) *Vanilla FL*, the standard federated learning framework without any privacy protection, serving as the upper-bound reference for model utility; and (2) *DP-enhanced FL*, which adds Gaussian noise to training images with a Differential-Privacy based method. Here the σ of Gaussian noise is 50.

B. Enhancement for Visual Privacy

Visual Obfuscation Effectiveness. Figure 5 demonstrates the visual obfuscation effects on CIFAR-10 images across three approaches. Vanilla FL uses raw training images with no obfuscation, leaving all object details clearly visible. FL+DP($\sigma = 50$) applies random noise to all pixels, yet fails to adequately obscure privacy-sensitive regions - object shapes like airplanes and animals remain distinguishable. In contrast, CSVAR's adaptive shuffling obfuscates images to show complete disruption of original object contours and textures, with no identifiable features remaining. This figure

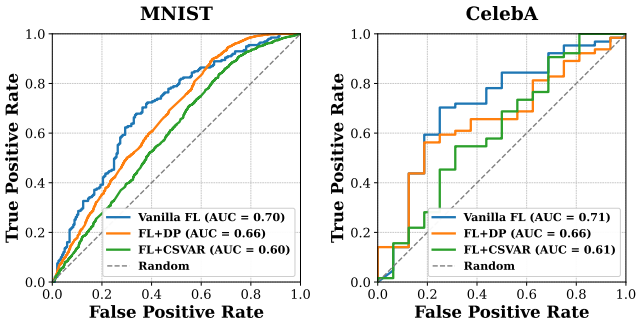

Fig. 7. ROC results of MIA with different protection methods. Low AUC means low attack Success Rate.

confirms CSVAR's superior perceptual ambiguity, particularly in preserving privacy for sensitive regions that DP fails to obfuscate adequately.

Resistance to GAN-based Reconstruction attacks. We evaluate GAN-based data reconstruction attacks [8] on models trained with different protection schemes. As shown in Figure 6, attacking models trained by vanilla FL yields nearly perfect reconstructions where all facial features from CelebA and number features from MNIST are clearly recoverable. While FL+DP-protected models produce visibly distorted outputs, salient facial and number features still remain decipherable to human eyes. With FL+CSVAR, the reconstructed images show only random noise patterns - no eyes, noses can be distinguished in facial images.

Mitigation of Membership inference attacks. Targeting the model weight after training, membership Inference attacks aim to determine whether a given data point belongs to the training dataset. Figure 7 demonstrates that membership inference attacks achieve high accuracy when targeting models trained with vanilla FL(the blue curve, AUC=0.7). For FL+DP, the attack performance decreases significantly(the orange curve, AUC=0.66), though some leakage remains detectable. CSVAR provides the strongest protection, where the attack AUC decreases to 0.6, close to the random guess curve.

C. Model Utility Preservation

We evaluate CSVAR's impact on model accuracy across three standard datasets (MNIST, CelebA, CIFAR-10) and three models (ResNet50, MobileNet, ShuffleNet), comparing against vanilla FL and FL+DP baselines. The results demonstrate CSVAR's ability to maintain model utility while providing strong visual privacy protection.

Accuracy Analysis. As demonstrated in Table I, CSVAR maintains model performance close to models trained with vanilla FL despite its strong privacy protections, in stark contrast to the substantial accuracy degradation with FL+DP. In general, on MNIST and CIFAR10 datasets, models trained with FL-CSVAR introduce negligible model utility loss(average 0.21%) compared to vanilla FL. Furthermore, across all evaluated models, FL+CSVAR demonstrates a consistent accuracy improvement over FL+DP, with an average accuracy increase of 3.23% on the CIFAR10 and a more pronounced 9.75% increase on the CelebA face dataset. Most notably, on CelebA, FL+CSVAR achieves 85.25% accuracy with ResNet50—demonstrating a mere 4% decrease from vanilla FL, while FL+DP exhibits a catastrophic 24.25% accuracy drop. This stark contrast underscores CSVAR's exceptional capability to preserve model utility while enhancing privacy.

		Resnet-50	ShuffleNet	MobileNet
MNIST	Vanilla FL	97.42%	96.88%	97.08%
	FL+DP	97.09%	95.44%	96.92%
	FL+CSVAR	97.39%	95.67%	97.03%
CIFAR10	Vanilla FL	84.24%	85.09%	85.18%
	FL+DP	80.49%	82.33%	80.03%
	FL+CSVAR	84.17%	83.71%	84.67%
CelebA	Vanilla FL	89.75%	85.50%	86.75%
	FL+DP	65.50%	77.25%	78.75%
	FL+CSVAR	85.25%	82.74%	82.75%

TABLE I
MODEL ACCURACY TRAINED WITH DIFFERENT PROTECTION METHODS.

VI. CONCLUSION

We propose CSVAR, a novel image shuffling framework to prevent visual privacy leakage from overfitting and the Client-side insecure private data transmission under Federated Learning. CSVAR adopts region-variance as the metric to measure a region's visual privacy sensitivity. CSVAR adaptively partitions each image region into smaller blocks with different granularity guided by region-variance, and then shuffles between blocks in both the spatial domains and chromatic channels. Experimental results show that CSVAR achieves a good trade-off between visual privacy protection against overfitting-induced privacy leakage and model utility.

REFERENCES

[1] S. Yue, Y. Deng, G. Wang, J. Ren, and Y. Zhang, "Federated offline reinforcement learning with proximal policy evaluation," *Chinese Journal of Electronics*, vol. 33, no. 6, pp. 1360–1372, 2024.

[2] Y. Deng, F. Lyu, J. Ren, H. Wu, Y. Zhou, Y. Zhang, and X. Shen, "Auction: Automated and quality-aware client selection framework for efficient federated learning," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 8, pp. 1996–2009, 2021.

[3] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The secret revealer: Generative model-inversion attacks against deep neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 250–258.

[4] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*. IEEE Computer Society, 2017, pp. 3–18.

[5] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients-how easy is it to break privacy in federated learning?" *Advances in neural information processing systems*, vol. 33, pp. 16 937–16 947, 2020.

[6] M. Abadi, A. Chu, I. J. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, E. R. Weippl, S. Katzenbeisser, C. Kruegel, A. C. Myers, and S. Halevi, Eds. ACM, 2016, pp. 308–318.

[7] Q. Li, Y. Zhang, J. Ren, Q. Li, and Y. Zhang, "You can use but cannot recognize: Preserving visual privacy in deep neural networks," in *31st Annual Network and Distributed System Security Symposium, NDSS 2024, San Diego, California, USA, February 26 - March 1, 2024*. The Internet Society, 2024.

[8] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the gan: information leakage from collaborative deep learning," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 2017, pp. 603–618.