Deep Reverse Attack on SIFT Features With a Coarse-to-Fine GAN Model

Xin Li, Guopu Zhu, *Senior Member, IEEE*, Shen Wang, Yicong Zhou, *Senior Member, IEEE*, and Xinpeng Zhang, *Senior Member, IEEE*

***Abstract*— Recently, it has been shown that adversaries can reconstruct images from SIFT features through reverse attacks. However, the images reconstructed by existing reverse attack methods suffer from information loss and are unable to sufficiently reveal the private contents of the original images. In this paper, a two-stage deep reverse attack model called Coarse-to-Fine Generative Adversarial Network (CFGAN) is proposed to more deeply explore the information in SIFT features and further demonstrate the risk of privacy leakage associated with SIFT features. Specifically, the proposed model consists of two sub-networks, namely coarse net and fine net. The coarse net is developed to restore coarse images using SIFT features, while the fine net is responsible for refining the coarse images to obtain better reconstruction results. To effectively leverage the information contained in SIFT features, an efficient fusion strategy based on the AdaIN operation is designed in the fine net. Additionally, we introduce a new loss function called sift loss that enhances the color fidelity of reconstructed images. Extensive experiments conducted on various datasets verify that the proposed CFGAN performs favorably against state-of-the-art methods. The reconstructed images exhibit better visual quality, less texture distortion, and higher color fidelity. Source code is available at https://github.com/HITLiXincodes/CFGAN.**

***Index Terms*— Data privacy, reverse attack, scale invariant feature transform (SIFT), generative adversarial network (GAN).**

I. Introduction

With the rapid growth of image information, image retrieval techniques have been widely applied in various fields [1], [2], [3]. The majority of related applications rely heavily on the local features extracted from the queried images [4], [5]. As one of the most popular local feature extraction and coding algorithms in computer vision, scale invariant feature transform (SIFT) [6] exhibits excellent matching performance when images undergo transformation or rotation [7]. In addition, SIFT is strongly robust to light changes and noise [8].

Due to the widespread usage of SIFT, the privacy and security issues linked with SIFT have drawn significant attention [9]. As a kind of local feature derived from images, the SIFT feature contains rich image content information [10]. It has been shown that reverse attacks can reconstruct original images from SIFT features [11]. Fig. 1 provides a brief illustration about the process of image content leakage caused by SIFT features. To fulfill image retrieval services, the computational capacity limitation of local devices requires users to transfer the queried SIFT features to remote service providers [12]. As a result, the image features shared with remote service providers can potentially be used to reconstruct the original images through reverse attacks [13].

In order to reveal the information within local features and evaluate the potential privacy risk caused by the abuse of local features, many image restoration methods [14], [15], [16], [17], [18], [19], [20], [21], [22] have been proposed to recover images from local features. The pioneering work [14] demonstrates that the reverse attack on feature descriptors is achievable under a range of conditions and configurations. Moreover, many different methods [15], [16], [17], [18] have shown that reverse attacks can be applied to a wide range of traditional image features, including SIFT, histogram of oriented gradient (HOG), and bag-of-words (BoW). The aforementioned approaches have demonstrated the possibility of reverse attack on local features. But there are significant disparities between the reconstruction results and the original images, and it remains challenging to determine the security risks caused by SIFT features.

With the development of neural network technology, recent works [19], [20], [21], [22], [23] have focused primarily on conducting image reverse attacks using convolutional neural networks (CNNs). Although these CNN-based models outperform traditional models, they still fail to sufficiently demonstrate the vulnerability of SIFT features. The images reconstructed by existing CNN-based models frequently exhibit severe edge artifacts or texture distortions, as well as information loss in terms of image details. Moreover, there is a noticeable disparity in color between the original images and the reconstructed images. To solve these problems, we propose an efficient image reconstruction model called Coarse-to-Fine Generative Adversarial Network (CFGAN). The two-stage

Manuscript received 1 November 2023; revised 8 January 2024; accepted 12 February 2024. Date of publication 20 February 2024; date of current version 3 July 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62172402 and Grant U22B2047; in part by the National Defense Basic Scientific Research Program of China under Grant JCKY2023603C043; in part by the Fundamental Research Funds for the Central Universities under Grant FRFCU5710011322 and Grant HIT.OCEF.2022050; in part by the Science and Technology Development Fund, Macau, under Grant 0049/2022/A1; in part by the University of Macau under Grant MYRG2022-00072-FST; and in part by the Key Research and Development Plan of Heilongjiang Province under Grant 2022ZX01C01. This article was recommended by Associate Editor T. Zhang. *(Corresponding author: Guopu Zhu.)*

Xin Li, Guopu Zhu, and Shen Wang are with the School of Cyberspace Science, Harbin Institute of Technology, Harbin 150001, China (e-mail: 23s003123@stu.hit.edu.cn; guopu.zhu@hit.edu.cn; shen.wang@hit.edu.cn).

Yicong Zhou is with the Department of Computer and Information Science, University of Macau, Macau, China (e-mail: yicongzhou@um.edu.mo).

Xinpeng Zhang is with the School of Computer Science, Fudan University, Shanghai 200433, China (e-mail: zhangxinpeng@fudan.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2024.3367808.

Digital Object Identifier 10.1109/TCSVT.2024.3367808

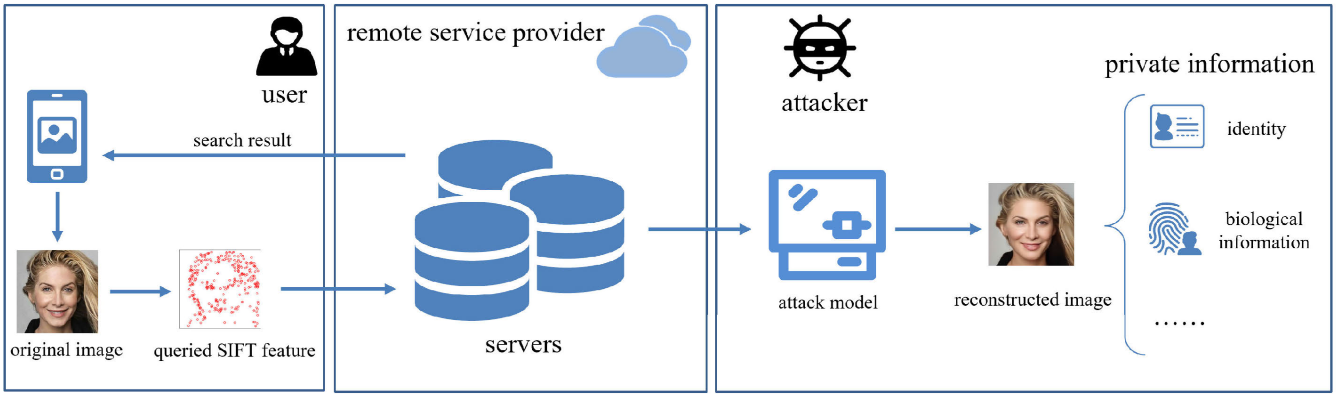

Fig. 1. Illustration of private image content leakage resulting from the misuse of SIFT features. While users upload queried features to remote service providers, attackers might abuse these features to reconstruct the original image, which leads to the leakage of private information. The attacker can be either a malicious hacker or the insider of an untrustworthy remote service provider.

reconstruction method used in CFGAN effectively alleviates the issues of edge artifacts and texture distortions, thereby significantly improving the quality of the reconstructed images. Additionally, a novel feature fusion strategy built on the AdaIN [24] operation is proposed in the fine net, which greatly optimizes the quality of the reconstructed image details. Furthermore, a new loss function named sift loss is designed for the purpose of enhancing the color accuracy of the reconstructed images.

In this paper, our major contributions are as follows:

- We propose a novel coarse-to-fine GAN-based model (CFGAN) for recovering latent images from SIFT features. The two-stage reconstruction method employed in the model not only significantly raises its capacity to recover image details but also effectively mitigates the issues of edge artifacts and texture distortions.
- A novel feature fusion strategy that uses the AdaIN operation is designed to perform multi-scale fusion between SIFT features and image features, which greatly improves the quality of reconstructed images.
- From the perspective of the similarity of SIFT feature descriptors, a new loss function called sift loss is proposed to improve the color consistency between the reconstructed images and the original images. The visual quality of the reconstructed images significantly benefits from the utility of sift loss.
- Extensive comparison experiments show that CFGAN outperforms state-of-the-art methods across a variety of datasets, while comprehensive ablation studies verify the effectiveness of each model component.

The rest of this paper is organized as follows: Section II provides a brief introduction to the SIFT algorithm, generative adversarial network (GAN), and existing reverse attack methods on image features. The process of reconstructing images from SIFT features by CFGAN is elaborated in Section III. Experimental setup and the analysis of experiment results are described in Section IV. In Section V, we draw a conclusion about this paper.

II. Related Work

In this section, we first briefly explain the SIFT feature extraction process. Then, the concept and structure of generative adversarial network (GAN) are described. At last, we introduce some important reverse attack methods on image features.

A. Extraction of SIFT Features

The extraction of SIFT features mainly involves four steps: establishment of the scale-space, accurate keypoint localization, orientation assignment, and construction of the local image descriptor.

1) Establishment of the Scale-Space: To establish a multi-level scale space, the Gaussian-blurred image $L(x, y, \sigma)$ can be calculated as

$$L(x, y, \sigma) = I(x, y) \otimes G(x, y, \sigma), \tag{1}$$

where $I(x, y)$ denotes the pixel value of image I at position (x, y), $\otimes$ represents convolutional operation, and $G(x, y, \sigma)$ is the Gaussian kernel at scale σ.

2) Accurate Keypoint Localization: The candidate set of feature points is obtained by comparing the values of 8 surrounding pixels at the same scale of the sampling points in the difference of Gaussian (DoG) pyramid and 9 pixels at adjacent scales.

3) Orientation Assignment: An orientation histogram is constructed by gathering the orientations within a localized region centered on SIFT keypoints. The maximum value in the orientation histogram is determined as dominant orientation.

4) Construction of the Local Image Descriptor: A 128-dimensional feature descriptor f is finally generated by calculating the gradient information of 8 directions in a 16×16 local area centered at the feature point.

B. Generative Adversarial Network

Generative adversarial network (GAN) is a machine learning framework first proposed by Goodfellow et al. [25], which consists of a generator and a discriminator. The generator is responsible for generating new data, while the discriminator evaluates the authenticity of the data in training samples. After the continuous optimization during the confrontation, the generator is capable of producing visually authentic images, while the discriminator is proficient in accurately distinguishing counterfeit images.

The exceptional performance of GAN networks has led to their widespread utilization in feature inversion tasks, resulting in the development of various GAN model variations [26], [27], [28], [29]. In order to improve the results of feature inversion, GleaD [26] focuses on addressing the issue of fairness between the generator and the discriminator in GAN networks. PadInv [27] uses the padding space of the generator to provide spatial information to the latent space, allowing the induction bias of pre-trained models to be suitably adjusted to each individual image. WaveGAN [28], a frequency-aware model for few-shot images, effectively synthesizes high-frequency signals with fine details. The model proposed in [29] develops a dual-path inpainting network with an inversion path and a feed-forward path, where the inversion path provides auxiliary information to help the feed-forward path. The success of these GAN models effectively demonstrates their superior capacity to perform feature inversion. Hence, selecting the GAN model as the tool for SIFT feature inversion work is a viable option.

C. Reverse Attacks on Image Features

The concept of reversing image features to obtain original images has been explored in recent years as image features play an increasingly important role. Various reverse attack methods have been proposed [14], [15], [16], [18] to recover original images from image features. Weinzaepfel et al. [14] first demonstrated the feasibility of restoring images from SIFT features. They used an exterior database of image patches for reconstructing original images. Angelo et al. [15] proposed an inverse optimization framework that is capable of recovering images only relying on the information carried by feature descriptors. Vondrick et al. [16] proposed a dictionary-learning-based approach to visualize HOG descriptors, which shows high transferability across a variety of different local features. Kato and Harada [18] showed that it is possible to recover some of the original image structures from sparse local descriptors in bag-of-words (BoW) representation.

With the prevalence of deep convolutional neural networks, many deep learning-based reverse attack methods [19], [20], [21], [22], [23] have been proposed. Mahendrand and Vedaldi [19] proposed a general framework based on neural networks to recover images, which significantly improves image recovery performance. The model proposed by Dosovitskiy and Brox [20] adopts an encoder-decoder structure to reconstruct images from local features. Furthermore, this model has been successfully applied to high-level features derived from convolutional neural networks. Pittaluga et al. [21] trained a cascade network with the structure of U-Net to reveal scenes from local features. The network effectively handles highly sparse and irregular 2D point distributions as well as inputs with missing point attributes. Wu and Zhou [22] improved image restoration performance by employing GANs architecture as the model backbone. Additionally, local binary pattern (LBP) features were utilized to compensate for the limitations of SIFT features in representing image spatial structures. A recent work by Pittaluga and Zhuang [23] proposed two novel inversion attacks, which show the feasibility of recovering the original image contents from after-processed image features. Although these methods achieve superior results to traditional approaches, they still have limitations in sufficiently revealing the information contained in SIFT features, and the quality of the reconstructed images is unsatisfactory.

III. Proposed Method

The overall architecture of CFGAN is first depicted in this section. Subsequently, each module in CFGAN is introduced in detail. Finally, we describe the loss functions used for CFGAN training.

A. Overview of CFGAN

As shown in Fig. 2, CFGAN consists of two sub-networks, namely coarse net and fine net. Both of the sub-networks follow the GAN model architecture, which includes a generator and a discriminator. However, each sub-network serves a different purpose. Specifically, the coarse net directly recovers image contents from the input SIFT feature S instead of converting the SIFT feature to LBP feature first as in [22]. The transformation from SIFT feature to LBP feature results in the loss of image information, and such information loss can be avoided if we directly restore the coarse image from the given SIFT features. For this reason, the coarse net is designed to directly restore image contents. However, the distribution of SIFT feature points is unbalanced, with a large concentration in certain image regions and an absence in others [30]. Consequently, the coarse image I_c reconstructed by the coarse net may exhibit distortion and deformation in terms of detailed texture, even a lack of content in some areas. Therefore, we design the fine net to further optimize the coarse image I_c. The fine net takes SIFT feature S and reconstructed coarse image I_c as inputs, aiming to improve the color accuracy and texture details of the coarse image. In the following, we will introduce the implementation details of these two sub-networks separately.

B. Coarse Net

The coarse net, depicted within the purple box in Fig. 2, is a convolutional encoder-decoder network. The encoder is in place for encoding the input SIFT features to obtain high-level feature maps, while the decoder performs decoding operations based on these feature maps to predict pixel values and reconstruct the target images. We adopt the U-Net architecture [31] for the structures of the encoder and decoder. Because it has been empirically demonstrated that the U-Net architecture is quite suitable for image generation [32]. The skip connection operation employed in U-Net accomplishes the direct transmission of low-level details from the encoder to the corresponding section of the decoder, thereby effectively improving the quality of reconstructed images.

First of all, it is worth mentioning the scale modification for the input SIFT feature. The number of SIFT feature points extracted from an image is small in comparison to the number of pixels. This means that a significant portion of the SIFT feature space is empty and dispensable for image recovery. To reduce the redundant information of input

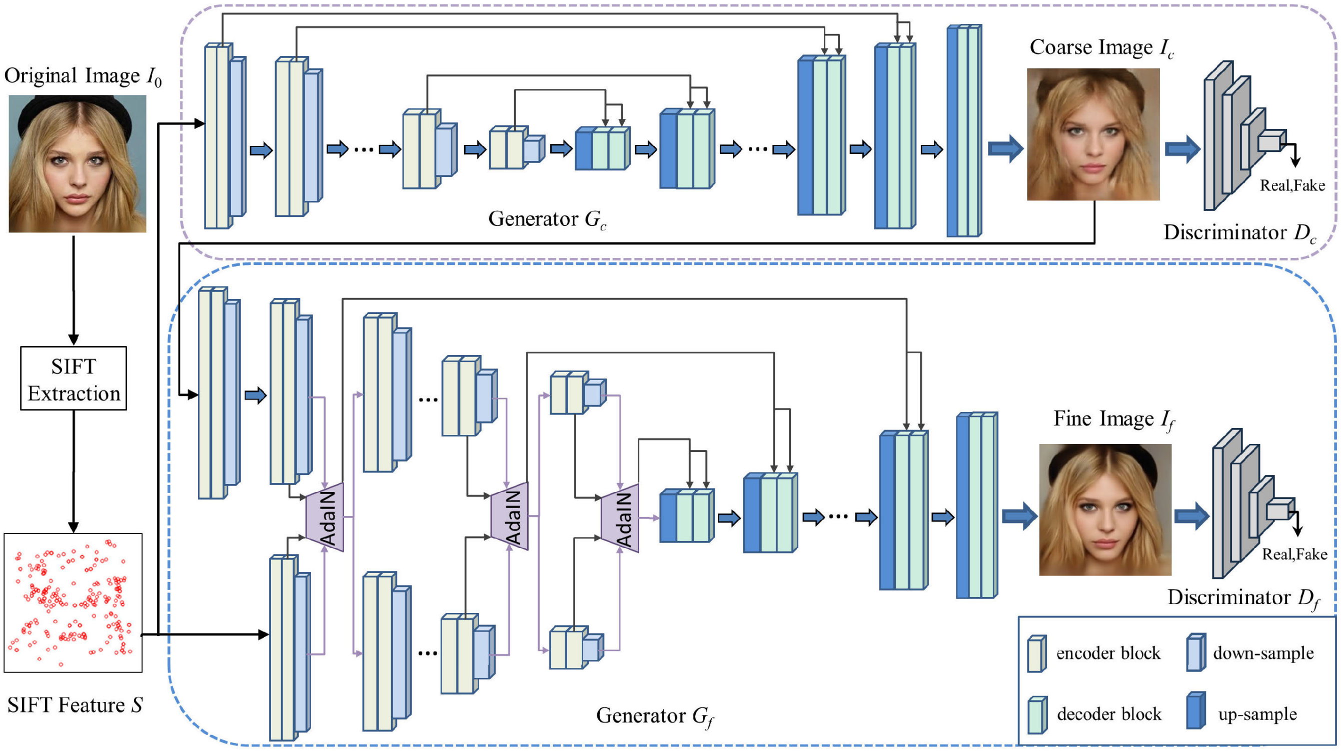

Fig. 2. Architecture of the proposed coarse-to-fine generative adversarial network (CFGAN). CFGAN is composed of two sub-networks: the coarse net, represented within the purple box, and the fine net, represented within the blue box. Both sub-networks consist of a generator and a discriminator.

SIFT feature, we propose a simple compression operation. Firstly, the SIFT feature is partitioned into small blocks of size 2 × 2. Subsequently, we compute the count of feature points within each block. If a block contains no feature point, its value is assigned as 0. If a block contains multiple feature points, a representative point is chosen randomly. This simple compression operation halves the input SIFT scale from 256 × 256 to 128 × 128. Experiments demonstrate that, despite a small loss in the number of feature points, the use of the proposed compression operation leads to a desirable reduction in computation and enhances recovery performance. Experimental results and analyses will be described in detail in the ablation study section.

The encoder of the coarse net is composed of a series of encoder blocks and down-sampling layers. Each encoder block with a down-sampling layer halves the scale of feature both in height and width. The size of the input SIFT feature is 128 × 128 with a total of 128 channels. After being processed by the encoder, the scale of the final high-level feature is 2 × 2 with 512 channels. The structure of the encoder block is displayed in Fig. 3. Two types of convolutional layers are utilized in the encoder block, one with a kernel size of 3 × 3, a stride of 1, a padding of 1, and the other employs a kernel size of 1 × 1 and a stride of 1. Each convolutional layer is followed by a batch normalization (BN) layer and a leaky rectified linear unit (LeakyReLU) layer. In order to alleviate the issue of gradient vanishing, we adopt the short-circuit connection introduced in ResNet [33] within the encoder block. Inspired by [34], we implement our down-sample operation using average pooling, which is beneficial to reduce the parameter count of the model.

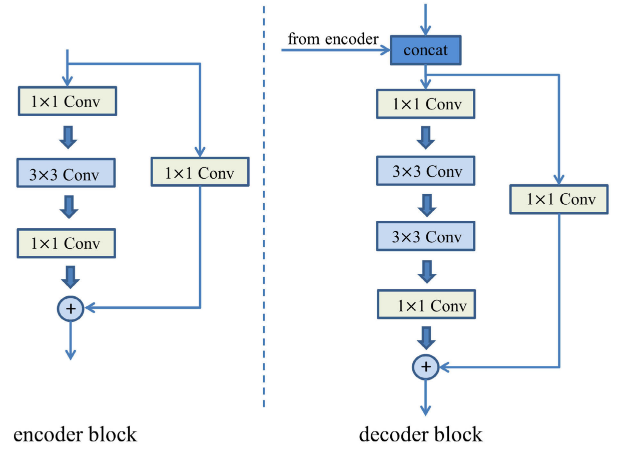

Fig. 3. Details of the encoder block and the decoder block. For convenience, batch normalization layers and activation functions are omitted.

The decoder of the coarse net consists of a range of up-sample layers and decoder blocks. The architecture of the decoder block is shown in Fig. 3. Compared with the encoder block, the decoder block uses an additional convolution layer with a kernel size of 3 × 3, along with a BN layer and a LeakyReLU layer. This is because we believe that incorporating a deeper network in the decoder part will help to better restore image content. To make the scale of feature maps double in both width and height, we employ the nearest-neighbor up-sampling layers.

The structure of the discriminator is comparatively simpler than that of the generator. Following the classic discriminator structure design, we use a series of convolutional layers to encode the input images. At last, a fully connected layer and the sigmoid function are applied to compute the probability values that represent the authenticity of images. The convolutional layers used in the discriminator are all identical, with a kernel size of 3 × 3, a stride of 1, and a padding of 1, except for the last one, whose kernel size is 4 × 4 and has no padding. The BN layer and the LeakyReLU layer are employed after each convolutional layer. The down-sampling operation in the discriminator is also performed by average pooling to help reduce the number of model parameters.

C. Fine Net

The architecture of the fine net is exhibited within the blue box of Fig. 2. The fine net comprises a generator and a discriminator, while the structures of the decoder and discriminator are consistent with the coarse net. However, in contrast to the coarse net, the encoder of the fine net is composed of two parts that share the same structure but different functions. The first part is responsible for encoding the input SIFT features, while the second part encodes the coarse images reconstructed by the coarse net. The encoder of the fine net is designed in this manner for two specific purposes. On one hand, there exist significant differences in terms of content and dimension between the SIFT features and the coarse images. The SIFT features include a series of key points, most of which appear at the edges of the objects within images. The coarse images contain abundant content information, which can clearly represent the color and texture of objects in images. On the other hand, the design of two independent encoder components facilitates a more flexible adjustment of the size and dimensions of both the SIFT features and the coarse images.

D. Fusion Strategy

The proposed fusion strategy plays an essential role in enhancing the capability of the fine net to improve the quality of image details. Inspired by [24] and [35], we employ the AdaIN operation as the fundamental component of our fusion strategy. The AdaIN operation itself, unlike widely used convolutional-based fusion modules [36], [37] and attention-based fusion modules [38], [39], does not introduce any extra parameter. It avoids the increase in model parameters caused by the addition of the fusion module. As can be seen from Fig. 2, we use the AdaIN operation to fuse the features obtained from each down-sample layer in the SIFT feature encoder and the coarse image encoder, as well as the features that are transmitted to the decoder. The AdaIN operation is formulated as follows:

$$\mathrm{AdaIN}(s, c) = \sigma(c)\left(\frac{s - \mu(s)}{\sigma(s)}\right) + \mu(c), \tag{2}$$

where s and c denote the feature maps encoded from the SIFT feature and the coarse image, respectively. Here, μ indicates the mean value of feature maps, and σ represents the standard deviation value of feature maps.

The purpose of integrating these two types of features at different spatial scales is to enhance the comprehensiveness of the fusion. As is widely acknowledged, the key points of SIFT features indicate the salient areas in the image. The multi-scale fusion technique effectively utilizes the indication information of SIFT features to reconstruct image details and rectify feature maps of coarse images at different scales, thereby achieving better detail recovery performance.

E. Loss Function

In order to optimize the CFGAN network, we use a combination of various loss functions. In particular, we design a new loss function named sift loss for our specific task that reverses SIFT features into images. Both the coarse net and the fine net share the same objective function.

1) Pixel Loss: We use the pixel-level loss function to constrain low-level information between the ground truth and the reconstructed image, which is consistent with many existing GAN models [22], [40]. The pixel loss is formulated as

$$\mathcal{L}_{pix} = \frac{1}{N}\sum_{i=1}^{N} ||I_0^i - I_g^i||_1, \tag{3}$$

where I_0 denotes the ground truth, I_g denotes the reconstructed image, and N represents the total number of pixels in the image. The notations for the ground truth and reconstructed image in the subsequent formulas remain consistent.

2) Perceptual Loss: The purpose of utilizing perceptual loss is to improve the visual quality of the reconstructed image. Specifically, we use a pre-trained VGG19 [41] to extract features about the ground truth and the reconstructed image. Therefore, the similarity between the ground truth and the reconstructed image can be calculated at the feature-level. The specific calculation formula for perceptual loss is defined as

$$\mathcal{L}_{per} = \sum_{l=1}^{L} ||\xi^l(I_0) - \xi^l(I_g)||_2, \tag{4}$$

where ξ^l indicates the l-th feature extraction layer in VGG, and L denotes the total number of feature extraction layers.

3) Adversarial Loss: The principle of GAN is that the generator desires to produce images that look real, while the discriminator makes an effort to distinguish fake images. Therefore, we adopt the adversarial function used in RaGAN [42], which can be defined as follows:

$$\mathcal{L}_{dis} = -\mathbb{E}_{I_0}[\log(D'(I_0))] - \mathbb{E}_{I_g}[\log(1 - D'(I_g))], \tag{5}$$

$$\mathcal{L}_{adv} = -\mathbb{E}_{I_g}[\log(D'(I_g))] - \mathbb{E}_{I_0}[\log(1 - D'(I_0))], \tag{6}$$

where

$$D'(I_g) = \mathrm{sigmoid}(D(I_g) - \mathbb{E}_{I_0}[D(I_0)], \tag{7}$$

$$D'(I_0) = \mathrm{sigmoid}(D(I_0) - \mathbb{E}_{I_g}[D(I_g)]. \tag{8}$$

4) Sift Loss: It is widely acknowledged that SIFT feature based image matching mainly depends on the similarity exhibited by feature descriptors. The feature descriptors are capable of capturing local gradient information surrounding feature points. So we assume that the similarity of SIFT features can also reflect the similarity of image contents. Based on this idea, we design a new loss function called sift loss to quantify

TABLE I

QUANTITATIVE COMPARISONS OF DIFFERENT METHODS OVER CELEBA-HQ [44], FFHQ [35], LSUN BIRD AND LSUN DINING ROOM [45] AMONG IVR [20], INV [21], SLI [22], GLEAD [26], EVDVAE [46] AND THE PROPOSED CFGAN. ↑ MEANS HIGHER IS BETTER, AND ↓ MEANS LOWER IS BETTER. THE BEST RESULTS ARE EMPHASIZED IN BOLD

Methods	CelebA-HQ			FFHQ			LSUN bird			LSUN dining room		
	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓
IVR [20]	0.4986	17.00	249.29	0.4393	15.69	346.76	0.3867	15.10	329.26	0.3989	15.68	301.64
INV [21]	0.6497	17.97	57.39	0.5816	16.41	141.14	0.6059	16.58	205.17	0.6288	17.58	126.07
SLI [22]	0.6869	18.77	61.92	0.5747	16.44	149.50	0.5716	16.13	215.10	0.5885	17.07	144.45
GleaD [26]	0.6401	18.31	62.67	0.5789	16.91	142.83	0.5214	15.62	225.68	0.5939	17.25	100.56
EVDVAE [46]	0.7138	19.99	50.90	0.6531	18.34	121.94	0.6213	17.41	220.75	0.6888	19.20	102.22
Ours	**0.7474**	**20.91**	**39.39**	**0.6888**	**19.15**	**100.79**	**0.6938**	**18.70**	**155.28**	**0.7311**	**20.07**	**65.47**

the distance between the original image and the reconstructed image by evaluating their disparities at the SIFT feature level. The sift loss can be formulated as

$$\mathcal{L}_{sift} = \frac{1}{M}\sum_{j=1}^{M} ||\mathcal{K}(I_0^j) - \mathcal{K}(I_g^j)||_1, \tag{9}$$

where $\mathcal{K}$ denotes the SIFT feature extraction operator included in the package of kornia [43] and M is the amount of SIFT feature points.

Finally, the generator of CFGAN is optimized by minimizing the function $\mathcal{L}_G$, while the discriminator of CFGAN is optimized by minimizing the function $\mathcal{L}_D$. The overall objective functions are defined as

$$\mathcal{L}_G = \lambda_{pix}\mathcal{L}_{pix} + \lambda_{per}\mathcal{L}_{per} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{sift}\mathcal{L}_{sift}, \tag{10}$$

$$\mathcal{L}_D = \mathcal{L}_{dis}, \tag{11}$$

where λ_{pix}, λ_{per}, λ_{adv}, and λ_{sift} show the parameters of the pixel loss, the perceptual loss, the adversarial loss, and the sift loss, respectively.

IV. EXPERIMENTS

A. Experimental Setup

1) Datasets: We utilize four publicly available datasets CelebA-HQ [44], FFHQ [35], LSUN dining room, and LSUN bird [45] for training and evaluating the proposed model. The CelebA-HQ contains 30,000 images, each of which features a human face as the main focus against a simple backdrop. We randomly divide them into training, validation, and testing sets in an 8:1:1 ratio. The FFHQ is a high-quality facial dataset, from which 3000 samples are chosen at random to evaluate the generalization ability of the model trained on the CelebA-HQ. The LSUN dining room consists of various complex living room images with no clear distinctions between the subjects and the background. A total of 50,000 images are randomly selected and divided into training, validation, and testing sets in an 8:1:1 ratio. The LSUN bird includes images with clean backgrounds as well as those with complicated backgrounds. We also randomly select 50,000 images in a distribution ratio consistent with the LSUN dining room. All the selected samples are resized to 256 × 256 pixels before training or testing.

2) Implementation Details: We implement our model using the PyTorch framework. All experiments are performed on two RTX 4090 GPUs. Meanwhile, we optimize our model by Adam [47], whose parameters are set as $\beta1 = 0.9$ and $\beta2 = 0.999$. The learning rate is set to 1×10^{-4}. We train the model with a batch size of 8, and the parameters trading off different terms in the loss functions are empirically fixed to be $\lambda_{pix} = 1$, $\lambda_{per} = 0.01$, $\lambda_{sift} = 0.01$, and $\lambda_{adv} = 0.02$. To improve the reconstruction performance and stabilize the training process of the model, we first train the coarse net to make sure that it can stably generate coarse images. Then, we employ the pre-trained coarse net to help train the fine net.

3) Evaluation Metrics: In order to objectively evaluate the quality of reconstructed images, we have selected three widely used image quality assessment metrics: peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [48], and fréchet inception distance (FID) [49].

B. Comparisons With Other Methods

In this part, we compare the image reconstruction performance of CFGAN with other state-of-the-art (SOTA) methods, including three SIFT-based image reconstruction methods IVR [20], INV [21], SLI [22] as well as two general image generation methods GleaD [26] and EVDVAE [46]. We re-implement the IVR and INV according to the descriptions in the papers, and we adopt the released codes of SLI, GleaD, and EVDVAE for training and testing. For fair comparisons, all models are trained and tested on the same datasets.

1) Quantitative Comparisons: The quantitative comparisons with other SOTA methods are presented in Table I. As can be seen, the CFGAN model outperforms all competing algorithms across four datasets, which strongly demonstrates its superior capability and remarkable effectiveness.

From the perspective of datasets, the proposed CFGAN model surpasses other approaches by a clear margin, especially in the LSUN bird and the LSUN dining room. As mentioned in the experimental setup, images in the LSUN dining room all contain a large number of objects, which implies that the SIFT features extracted from these images will contain plentiful parameters. It is undoubtedly more challenging to restore intricate object details in complex scenes based on

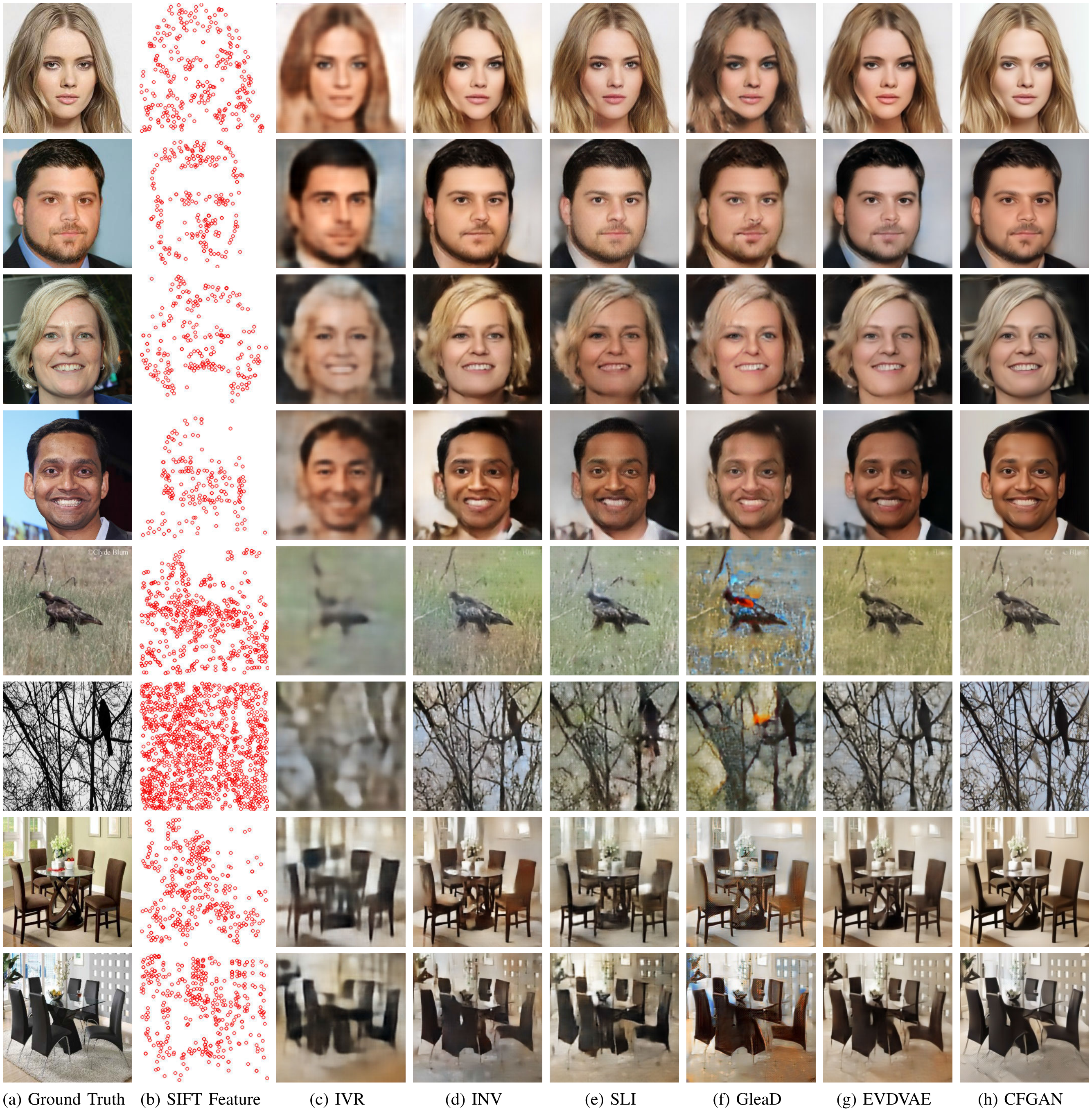
(a) Ground Truth (b) SIFT Feature (c) IVR (d) INV (e) SLI (f) GleaD (g) EVDVAE (h) CFGAN

Fig. 4. Visual comparisons of different methods. From top to bottom, the images in each pair of rows are sourced from CelebA-HQ [44], FFHQ [35], LSUN bird, and LSUN dining room [45], respectively. Within each row, the sequence is as follows: ground truth, SIFT feature, results reconstructed by IVR [20], INV [21], SLI [22], GleaD [26], EVDVAE [46] and CFGAN.

large amounts of SIFT features compared to only restoring the main content of images with simplistic backgrounds. Therefore, the experimental results in the LSUN dining room demonstrate that CFGAN exhibits superior capabilities in restoring complex image details compared to the competing models. In terms of the LSUN bird, it contains both simple and complex images, which means that the number of input SIFT feature points varies significantly, ranging from hundreds to thousands. The results obtained from the LSUN bird indicate that CFGAN shows exceptional adaptability to diverse input scenarios, thereby enhancing its practical applicability.

Considering the experimental results obtained on these four datasets, the proposed CFGAN is capable of conducting reverse attacks on SIFT features in various situations. Even if the original image exhibits a complicated scene, the reverse attack might still be able to recover much of the scene content, including tiny details. This illustrates that the reverse attack on SIFT features can be achieved regardless of the type of the original image, which further demonstrates the potential image privacy leakages caused by the SIFT features.

From the perspective of evaluation metrics, the proposed CFGAN model shows the most significant improvement in

TABLE II

COMPARISONS OF THE MODEL PARAMETERS, FLOPS, AND INFERENCE TIME ON DIFFERENT METHODS

Methods	Params	FLOPs	Inference time
IVR [20]	93M	25G	0.085s
INV [21]	88M	150G	0.044s
SLI [22]	78M	96G	0.129s
GleaD [26]	147M	121G	0.090s
EVDVAE [46]	198M	304G	0.136s
CFGAN	153M	148G	0.135s

TABLE III

VERIFY THE EFFECTIVENESS OF COMPRESSION OPERATION. THE BEST RESULTS ARE EMPHASIZED IN BOLD

Models	CelebA-HQ		
	SSIM↑	PSNR↑	FID↓
w/o compression	0.7384	20.59	44.64
w compression	**0.7402**	**20.81**	**43.38**

terms of the FID metric, illustrating that the distribution of reconstructed images closely approximates the original image distribution. Clear improvements can also be observed in SSIM and PSNR across multiple datasets, which reflect the enhancement in the visual quality of the images. It indicates that the images restored by CFGAN are better aligned with the original images in terms of color, brightness, and content. Moreover, the images reconstructed by CFGAN exhibit high quality. It shows that the reverse attack on SIFT features is capable of recovering the majority of the original image content, including the information that is associated with user privacy. This emphasizes the existence of the risk of image privacy leakage associated with SIFT features.

Supplementary information about each compared method is listed in Table II, which includes the model parameters, the FLOPs, and the inference time required for a single image. Although IVR and INV require fewer model parameters and require less inference time than CFGAN, the quality of the images reconstructed by these two methods is far from satisfactory. In addition, CFGAN holds a slightly higher number of model parameters than GleaD and a slightly slower inference speed than SLI, but CFGAN shows better reconstruction performance than these two models. Furthermore, EVDVAE is more complex than CFGAN in terms of model architecture, but its reconstruction performance is inferior to CFGAN, which also illustrates the superiority of CFGAN.

2) Qualitative Results: The reconstructed samples by competitive methods and CFGAN are presented in Fig. 4, which enable us to make a more intuitive comparison of the reconstruction performance displayed by these models. The IVR model can reconstruct the overall contour information of the subjects. However, there are noticeable areas of blurriness and a significant lack of details in the reconstructed images. The images generated by the INV model and the SLI model show a high degree of similarity in terms of content to the original images. But noticeable distortions still exist in edge positions and areas with significant variations in brightness. Additionally, these images exhibit an issue of color degradation. Both the GleaD model and the EVDVAE model have made considerable advancements in processing image details, but there are still limitations in terms of image color. Compared to the aforementioned models, the images reconstructed by the proposed CFGAN model exhibit minimal distortion and deformation in terms of object edges and textures, which greatly enhances their visual coherence with the original images. Besides, the CFGAN model shows a commendable ability to accurately reconstruct tiny objects within intricate scenes, indicating that CFGAN can deal with more challenging image reconstruction tasks. Furthermore, the images reconstructed by the CFGAN model hold more accurate color consistency with the original images, which significantly enhances their visual quality. Visual comparisons in Fig. 4 show that the images reconstructed by the proposed model are quite similar to the original images, involving the reconstruction of small objects in a living room scenario or significant facial features. It validates the feasibility of restoring the private content of the original image from SIFT features via reverse attacks. In other words, if the SIFT features extracted from the original images are exposed to an untrusted third party, it may cause serious privacy disclosure.

C. Ablation Study

In this part, we independently examine each model component to ensure its validity. The experimental results and analysis of the compression operation on input SIFT features are presented first. In order to confirm the effectiveness of the proposed network architecture, we then look at the individual contribution of each structure component in the absence of sift loss. Finally, we incorporate the sift loss into each ablation model to evaluate the impact of the sift loss function on model reconstruction capability.

1) Study of the Compression Operation: To validate the effectiveness of the proposed compression operation on input SIFT features, we designed two comparative models based on the CFGAN architecture. The first model, marked as “w/o compression”, does not use the compression operation. The second model, marked as “w compression”, employs the proposed compression operation. Experiment results are listed in Table III, which demonstrate the effectiveness of the compression operation across all three evaluation metrics. Additionally, compressing the scale of input SIFT features allows us to reduce the number of network layers in the CFGAN encoder, which lowers the model parameters and speeds up computation. We consider that the compression operation effectively reduces a large amount of redundant information in the original SIFT features, which mitigates the sparsity issue inherent to the SIFT features and directs the model to concentrate on areas with more abundant information.

2) Study of the Two-Stage Reconstruction Method: CFGAN utilities a two-stage reconstruction method to recover the original images. In this part, we conduct a set of experiments to

TABLE IV
STUDY OF THE PROPOSED TWO-STAGE RECONSTRUCTION METHOD. THE BEST RESULTS ARE EMPHASIZED IN BOLD

Models	CelebA-HQ			FFHQ			LSUN bird			LSUN dining room		
	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓
w/o fine net	0.7286	20.45	45.95	0.6691	18.79	110.26	0.6576	17.93	177.96	0.7031	19.65	86.03
w fine net	**0.7402**	**20.81**	**43.38**	**0.6812**	**19.10**	**103.41**	**0.6781**	**18.44**	**166.98**	**0.7174**	**20.02**	**77.13**

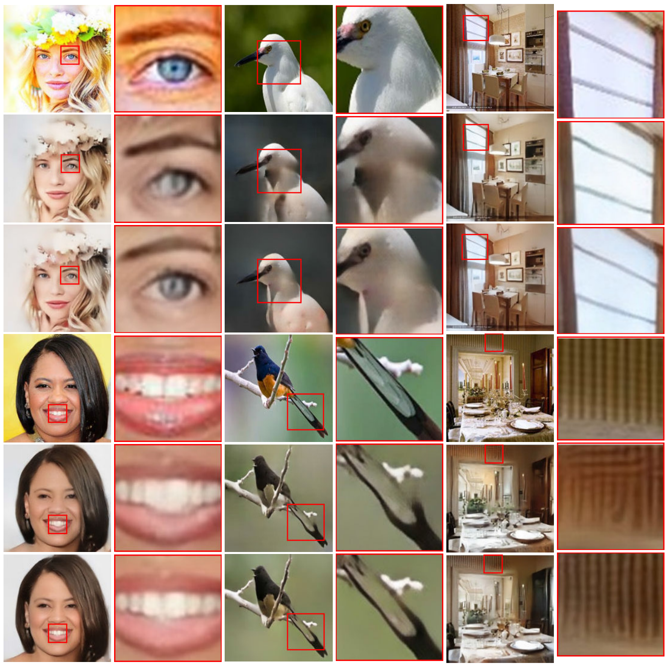

Fig. 5. Visual results on the ablation study of the proposed two-stage reconstruction method. The images are organized into two groups, with the first three rows comprising one group and the last three rows forming the other. In each group, images within the first row are ground truths, images reconstructed by the model “w/o fine net” are shown in the second row, and images reconstructed by the model “w fine net” are shown in the third row.

demonstrate the effect of this method. Specifically, we design two ablation models with different architectures. Model 1, denoted by “w/o fine net”, removes the whole fine net. Model 2, denoted by “w fine net”, retains the complete fine net structure. From Table IV, it can be observed that the incorporation of fine net significantly improves the model performance, which strongly demonstrates the effectiveness of the proposed two-stage reconstruction method. Some reconstruction examples are provided in Fig. 5. It is clearly shown that the coarse net is able to reconstruct the basic content of images, while the fine net effectively enhances image details, which leads to fewer texture distortions and artifacts. This further proves that adopting the two-stage reconstruction method can well improve the quality of reconstructed images.

3) Study of the Proposed Fusion Strategy: The proposed fusion strategy is the most critical part of fine net, which is designed to facilitate multi-level fusion of SIFT features and coarse images. To verify the effectiveness of the fusion strategy, we carry out a series of ablation experiments in this part. The first model, denoted by “w/o fusion”, replaces the AdaIN based fusion module with a simple concatenation operation that combines the SIFT feature and the coarse image at the input of fine net. The second model, denoted by “ w fusion”, employs the proposed fusion strategy. Quantitative results are provided in Table V. The importance of the proposed fusion strategy is demonstrated by the fact that applying the feature fusion strategy can significantly improve model performance.

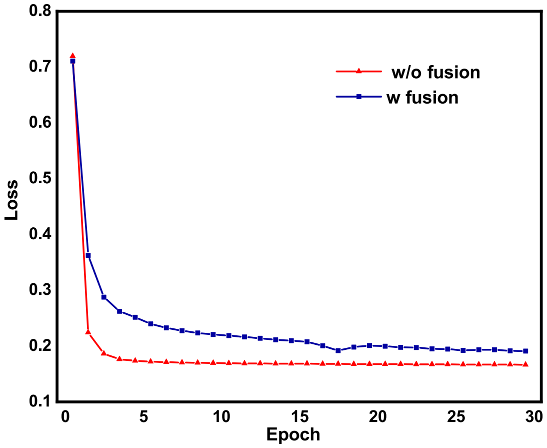

Fig. 6. Loss changes of the models “w/o fusion” and “w fusion” in initial thirty epochs during the training process.

The analysis of the training losses may provide insight into the effect of the fusion strategy. As is shown in Fig. 6, the training loss about the “w/o fusion” model descends quite rapidly and tends to stabilize in the first few epochs, while the training loss about the “w fusion” model declines gently. It indicates that the “w/o fusion” model appears to converge quickly, and the SIFT features input in the fine net are disregarded and make no contribution to the optimization of coarse images. We consider that the fine net without the fusion module plays the same role as the coarse net in optimizing the objective function. Because the incomplete fine net and the coarse net share the same network structures. The proposed fusion strategy is able to reorganize the SIFT features and coarse image features across different scales, thereby altering the direction of information flow and improving model performance. It further substantiates the indispensability of the fusion strategy in fine net as it effectively enhances the image reconstruction capability of the proposed model.

4) Effectiveness of the Sift Loss: In order to test the effectiveness of the sift loss and compare its effect with the style loss proposed in [22], we conduct an ablation study by designing two sets of comparative experiments. The first set includes three models, where model one is marked as “$\mathcal{L}_{pix+per+adv}$(c)”, model two is marked as “w $\mathcal{L}_{style}$(c)” and

TABLE V

Study of the Proposed Fusion Strategy. The Best Results Are Emphasized in Bold

Models	CelebA-HQ			FFHQ			LSUN bird			LSUN dining room		
	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓
w/o fusion	0.7123	20.22	**39.16**	0.6542	18.57	**93.95**	0.6583	17.92	169.04	0.7150	19.81	**71.92**
w fusion	**0.7402**	**20.81**	43.38	**0.6812**	**19.10**	103.41	**0.6781**	**18.44**	**166.98**	**0.7174**	**20.02**	77.13

TABLE VI

Study of the Loss Function. The Best Results Are Emphasized in Bold

Models	CelebA-HQ			FFHQ			LSUN bird			LSUN dining room		
	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓
$\mathcal{L}_{pix+per+adv}$(c)	0.7286	20.45	45.95	0.6691	18.79	110.26	0.6576	17.93	177.96	0.7031	19.65	86.03
w $\mathcal{L}_{style}$(c)	0.7150	20.23	51.61	0.6564	18.62	121.87	0.6231	17.57	207.97	0.6999	19.54	95.00
w $\mathcal{L}_{sift}$(c)	0.7462	20.79	**39.35**	0.6888	19.08	**99.63**	0.6805	18.34	161.85	0.7255	19.86	68.08
$\mathcal{L}_{pix+per+adv}$(c&f)	0.7402	20.81	43.38	0.6812	19.10	103.41	0.6781	18.44	166.98	0.7174	20.02	71.92
w $\mathcal{L}_{style}$(c&f)	0.7326	20.63	46.69	0.6730	18.87	114.74	0.6522	17.88	191.15	0.7161	19.90	85.24
w $\mathcal{L}_{sift}$(c&f)	**0.7474**	**20.91**	39.39	**0.6890**	**19.15**	100.79	**0.6938**	**18.70**	**155.28**	**0.7311**	**20.07**	**65.47**

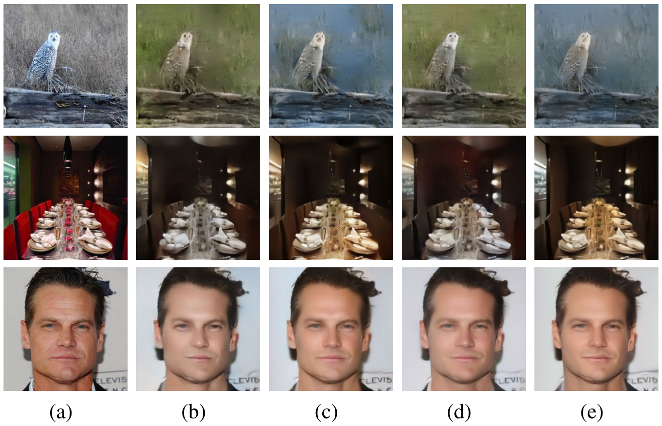

Fig. 7. Qualitative results of the ablation study that tests the effectiveness of sift loss. (a) Original images. (b) Results from the model $\mathcal{L}_{pix+per+adv}$(c). (c) Results from the model w $\mathcal{L}_{sift}$(c). (d) Results from the model $\mathcal{L}_{pix+per+adv}$(c&f). (e) Results from the model w $\mathcal{L}_{sift}$(c&f). Please zoom in for better observation.

model three is marked as “w $\mathcal{L}_{sift}$(c)”. All models employ the coarse net as the network framework. The distinction is that model one lacks style loss and sift loss, whereas model two includes style loss and model three includes sift loss. The second set consists of models four, five, and six, which are labeled as “$\mathcal{L}_{pix+per+adv}$(c&f)”, “w $\mathcal{L}_{style}$(c&f)” and “w $\mathcal{L}_{sift}$(c&f)”, respectively. CFGAN serves as the network architecture in all three models, with the only difference being that model four lacks the style loss and sift loss, model five includes the style loss, and model six includes the sift loss. The quantitative results are given in Table VI. We can notice that the incorporation of sift loss leads to a significant enhancement in the performance of both the coarse net and CFGAN. However, adding style loss will result in a clear decline in the performance of both the coarse net and CFGAN, which suggests that style loss is not suitable for CFGAN.

Notably, the most significant improvement is observed in terms of the SSIM metric, which indicates that the employment of sift loss enables the model to reconstruct images with better visual quality. As shown in Fig. 7, the qualitative results also serve to verify the effectiveness of sift loss. The sift loss not only helps the model to better reconstruct image details but also significantly improves the color fidelity of reconstructed images. The enhancement of color fidelity in the reconstructed images brings them closer to the original images in terms of visual perception. We consider that the sift loss can modify the color of reconstructed images by quantifying the disparity between the SIFT feature descriptors extracted from the reconstructed images and original images, which is of benefit to improving color fidelity in the reconstructed images.

V. Conclusion

In this work, we have proposed a novel deep generation model called Coarse-to-Fine Generative Adversarial Network (CFGAN) to reveal the latent information in SIFT features and assess the privacy risk that may arise from the leakage of SIFT features. CFGAN provides a two-stage reconstruction approach to recover the original images. Specifically, the first stage concentrates on the reconstruction of primary image content, while the second stage further enhances image details. To better utilize the SIFT feature information, we have introduced an efficient fusion strategy that employs the AdaIN operation. Furthermore, a new loss function named sift loss has been designed to improve the color fidelity of reconstructed images. Extensive experiments have demonstrated that the proposed CFGAN model outperforms the existing methods on four benchmark datasets.

However, there are certain limitations with the CFGAN model, primarily with regard to its generalizability and capacity for handling images with complex content. Improvements will be made in our future work to overcome these limitations. First, in order to enhance generalization ability, we can explore various mappings between samples through transfer learning. In addition, designing more effective loss functions may also be an alternative way to enhance the capacity of the model to reconstruct images with more details.

REFERENCES

[1] S. Jain, K. Pulaparthi, and C. Fulara, "Content based image retrieval," *Int. J. Adv. Eng. Glob. Technol.*, vol. 3, no. 10, pp. 1251–1258, 2015.

[2] X. Chai, Y. Wang, Z. Gan, X. Chen, and Y. Zhang, "Preserving privacy while revealing thumbnail for content-based encrypted image retrieval in the cloud," *Inf. Sci.*, vol. 604, pp. 115–141, Aug. 2022.

[3] W. Nie, Y. Zhao, J. Nie, A.-A. Liu, and S. Zhao, "CLN: Cross-domain learning network for 2D image-based 3D shape retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 992–1005, Mar. 2022.

[4] N. Keisham and A. Neelima, "Efficient content-based image retrieval using deep search and rescue algorithm," *Soft Comput.*, vol. 26, no. 4, pp. 1597–1616, Feb. 2022.

[5] W. Xiong, Z. Xiong, Y. Cui, L. Huang, and R. Yang, "An interpretable fusion Siamese network for multi-modality remote sensing ship image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2696–2712, Jun. 2022.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[7] Y. Xiang, F. Wang, and H. You, "OS-SIFT: A robust SIFT-like algorithm for high-resolution optical-to-SAR image registration in suburban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3078–3090, Jun. 2018.

[8] Y. Shi, Z. Lv, N. Bi, and C. Zhang, "An improved SIFT algorithm for robust emotion recognition under various face poses and illuminations," *Neural Comput. Appl.*, vol. 32, no. 13, pp. 9267–9281, Jul. 2020.

[9] X. Bi, C. Shuai, B. Liu, B. Xiao, W. Li, and X. Gao, "Privacy-preserving color image feature extraction by quaternion discrete orthogonal moments," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 1655–1668, 2022.

[10] X. Liu, X. Zhao, Z. Xia, Q. Feng, P. Yu, and J. Weng, "Secure outsourced SIFT: Accurate and efficient privacy-preserving image SIFT feature extraction," *IEEE Trans. Image Process.*, vol. 32, pp. 4635–4648, 2023.

[11] A. Tonge and C. Caragea, "Image privacy prediction using deep neural networks," *ACM Trans. Web*, vol. 14, no. 2, pp. 1–32, May 2020.

[12] S. Li and X. Li, "Secure image retrieval based on deep learning in cloud computing," in *Proc. 8th Int. Conf. Multimedia Image Process.*, Apr. 2023, pp. 40–46.

[13] P. Sun, "Security and privacy protection in cloud computing: Discussions and challenges," *J. Netw. Comput. Appl.*, vol. 160, Jun. 2020, Art. no. 102642.

[14] P. Weinzaepfel, H. Jégou, and P. Pérez, "Reconstructing an image from its local descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jun. 2011, pp. 337–344.

[15] E. d'Angelo, A. Alahi, and P. Vandergheynst, "Beyond bits: Reconstructing images from local binary descriptors," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 935–938.

[16] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba, "HOGgles: Visualizing object detection features," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1–8.

[17] A. Desolneux and A. Leclaire, "Stochastic image reconstruction from local histograms of gradient orientation," in *Proc. Int. Conf. Scale Space Variational Methods Comput. Vis.*, 2017, pp. 133–145.

[18] H. Kato and T. Harada, "Image reconstruction from bag-of-visual-words," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 955–962.

[19] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5188–5196.

[20] A. Dosovitskiy and T. Brox, "Inverting visual representations with convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jun. 2016, pp. 4829–4837.

[21] F. Pittaluga, S. J. Koppal, S. B. Kang, and S. N. Sinha, "Revealing scenes by inverting structure from motion reconstructions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 145–154.

[22] H. Wu and J. Zhou, "Privacy leakage of SIFT features via deep generative model based image reconstruction," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2973–2985, 2021.

[23] F. Pittaluga and B. Zhuang, "LDP-Feat: Image features with local differential privacy," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 17580–17590.

[24] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1501–1510.

[25] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. Neural Info. Process. Syst.*, 2014, pp. 2672–2680.

[26] Q. Bai, C. Yang, Y. Xu, X. Liu, Y. Yang, and Y. Shen, "GLeaD: Improving GANs with a generator-leading task," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12094–12104.

[27] M. Yang, Z. Wang, Z. Chi, and W. Feng, "WaveGAN: Frequency-aware GAN for high-fidelity few-shot image generation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 1–17.

[28] Q. Bai, Y. Xu, J. Zhu, W. Xia, Y. Yang, and Y. Shen, "High-fidelity GAN inversion with padding space," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 36–53.

[29] W. Wang, L. Niu, J. Zhang, X. Yang, and L. Zhang, "Dual-path image inpainting with auxiliary GAN inversion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2022, pp. 11421–11430.

[30] L. Wang, C. Lan, B. Wu, T. Gao, Z. Wei, and F. Yao, "A method for detecting feature-sparse regions and matching enhancement," *Remote Sens.*, vol. 14, no. 24, p. 6214, Dec. 2022.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[32] Q. Luo, H. Li, Z. Chen, and J. Li, "ADD-UNet: An adjacent dual-decoder UNet for SAR-to-optical translation," *Remote Sens.*, vol. 15, no. 12, p. 3125, Jun. 2023.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] R. Child, "Very deep VAEs generalize autoregressive models and can outperform them on images," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–17.

[35] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jun. 2019, pp. 4401–4410.

[36] W. Zhou, Q. Guo, J. Lei, L. Yu, and J. Hwang, "ECFFNet: Effective and consistent feature fusion network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1224–1235, Mar. 2022.

[37] G. Li, J. Wu, C. Deng, and Z. Chen, "Parallel multi-fusion convolutional neural networks based fault diagnosis of rotating machinery under noisy environments," *ISA Trans.*, vol. 128, pp. 545–555, Sep. 2022.

[38] T. Zhou, S. Ruan, P. Vera, and S. Canu, "A tri-attention fusion guided multi-modal segmentation network," *Pattern Recognit.*, vol. 124, Apr. 2022, Art. no. 108417.

[39] X. Wang et al., "Self-paced feature attention fusion network for concealed object detection in millimeter-wave image," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 224–239, Jan. 2022.

[40] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.* Oxford, U.K.: Computational and Biological Learning Society, 2015, pp. 1–14.

[42] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard GAN," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–26.

[43] E. Riba, D. Mishkin, D. Ponsa, E. Rublee, and G. Bradski, "Kornia: An open source differentiable computer vision library for PyTorch," 2019, *arXiv:1910.02190*.

[44] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–26.

[45] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop," 2015, *arXiv:1506.03365*.

[46] L. Hazami, R. Mama, and R. Thurairatnam, "Efficient-VDVAE: Less is more," 2022, *arXiv:2203.13751*.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

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

[49] 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," in *Proc. Neural Info. Process. Syst.*, 2017, pp. 6629–6640.

Xin Li received the B.S. degree in computer science from Harbin Institute of Technology, China, in 2023, where he is currently pursuing the M.S. degree with the School of Cyberspace Science. His primary research interests include multimedia security and image processing.

Guopu Zhu (Senior Member, IEEE) received the B.S. degree in transportation from Jilin University, China, in 2002, and the M.S. and Ph.D. degrees in control science and engineering from Harbin Institute of Technology, China, in 2004 and 2007, respectively. He is currently a Professor with Harbin Institute of Technology. He has authored or coauthored more than 60 articles in peer-reviewed international journals. His main research areas are multimedia security, image processing, and control theory. He serves as an Associate Editor for several journals, including IEEE Transactions on Cybernetics, IEEE Systems Journal, *Journal of Information Security and Applications*, and *Electronics Letters*.

Shen Wang received the B.S. and M.E. degrees in electrical engineering and information technology from the TUD-Dresden University of Technology, Dresden, Germany, in 2003 and 2007, respectively, and the Ph.D. degree in computer science and technology from Harbin Institute of Technology, Harbin, China, in 2012. Currently, he is a Professor with the School of Cyberspace Science, Harbin Institute of Technology. His current research interests include the adversarial attack and defense based on machine learning, image disguise, digital forensics, and quantum information processing.

Yicong Zhou (Senior Member, IEEE) received the B.S. degree in electrical engineering from Hunan University, Changsha, China, and the M.S. and Ph.D. degrees in electrical engineering from Tufts University, Medford, MA, USA.

He is currently a Professor with the Department of Computer and Information Science, University of Macau, Macau, China. His research interests include image processing, computer vision, machine learning, and multimedia security.

Dr. Zhou is a fellow of the Society of Photo-Optical Instrumentation Engineers (SPIE) and was recognized as one of "Highly Cited Researchers" in 2020, 2021, and 2023. He serves as an Associate Editor for IEEE Transactions on Cybernetics, IEEE Transactions on Neural Networks and Learning Systems, and IEEE Transactions on Geoscience and Remote Sensing.

Xinpeng Zhang (Senior Member, IEEE) received the B.S. degree in computational mathematics from Jilin University, China, in 1995, and the M.E. and Ph.D. degrees in communication and information system from Shanghai University, China, in 2001 and 2004, respectively. Since 2004, he has been a Faculty Member of the School of Communication and Information Engineering, Shanghai University, where he is currently a Professor. He is also a Faculty Member of the School of Computer Science, Fudan University. He was with The State University of New York at Binghamton as a Visiting Scholar from 2010 to 2011 and also with Konstanz University as an experienced Researcher, sponsored by the Alexander von Humboldt Foundation from 2011 to 2012. His research interests include multimedia security, AI security, and image processing. He has published over 300 articles in these areas. He was an Associate Editor of IEEE Transactions on Information Forensics and Security from 2014 to 2017.