sciendo

JAISCR, 2026, Vol. 16, No. 1, pp. 39 – 53
10.2478/jaiscr-2026-0002

S^{3}DIFF: SEMANTIC FUSION AND STRUCTURE-GUIDED GLOBAL GENERATION FROM A SINGLE IMAGE WITH DIFFUSION MODELS

Xianjie Zhang, Yusen Zhang, Yujie He*, Min Li*

Xi'an High-Tech Research Institute,
Xi'an, Shaanxi, China

E-mail: ksy5201314@163.com, proflimin@163.com

Submitted: 6th June 2025; Accepted: 22th August 2025

Abstract

Single-image generation models generate high-quality and diverse images by learning the internal distribution of patches within a single image, addressing the issue of data scarcity and attracting increasing attention. However, existing methods are unsatisfactory when dealing with images with global structures, such as animal images. To address this issue, we propose **S**emantic fusion and **S**tructure-guided global generation from a **S**ingle image with **Diff**usion models (**S^{3}Diff**). Specifically, during training, we employ a semantic extractor to extract high-level semantic features from training images and use the proposed semantic fusion block to fuse semantic features with image features, enhancing the model's understanding of image semantics and improving the quality of the generated images. During sampling, we apply manifold constrained gradient based on image structure to enforce the generation path to regress to the manifold of the original image, preserving reasonable global structures. Extensive experiments on public datasets demonstrate the thorough exploration of hyperparameters and the rationality of key designs, with quantitative and qualitative comparisons against baseline methods and validating that our proposed method preserves reasonable semantic and structural relationships, can generate high-quality and diverse images, significantly improving the model's global generation capabilities.

Keywords: diffusion models, single-image generation models, global generation

1 Introduction

Image generation has evolved from early traditional methods to deep learning-based approaches, making it one of the most attractive research directions in the field of computer vision [18, 19, 20, 21, 22, 23, 24, 25, 26]. Compared to early image generation methods, deep learning-based methods can handle massive amounts of training data and learn more refined and complex feature representations, significantly improving the quality and diversity of generated images. However, deep learning-based image generation methods typically require large amounts of training data, which is challenging to obtain in data-scarce domains, such as the medical field [27].

Therefore, training a deep generative model using a single image has attracted more and more attention. The pioneer of this method is SinGAN [1], which can generate diverse images similar to training images by learning the internal distribution

of patches within a single image. Subsequently, single-image generation models based on Generative Adversarial Networks (GANs) have sprung up [2, 3, 5, 28]. With the increasingly superior performance of diffusion models in image generation tasks, some researchers have also used diffusion models to train single-image generation models. Inspired by SinGAN, SinDDM [9] treats scale as a training parameter, using a single natural image to train a diffusion model, and gradually transitions from small to large scales during sampling to generate images. SinDiffusion [7] and SinFusion [8] propose using a single natural image to train a diffusion model at a single scale, removing attention blocks to ensure the model can learn the internal distribution of patches within a single image.

Although the methods above perform well in generating single images, significant improvements are still needed when handling images with global structures, such as animal images. Some researchers have explored the preservation of the global structures of images. ExSinGAN [4], based on multi-scale GAN, addresses complex image generation from three aspects: structure, semantics, and texture. However, the multi-scale structures may accumulate errors, resulting in unsatisfactory generation images with some artifacts. One-shot GAN [29], based on single-scale GAN, proposes a two-branch discriminator, enabling the generator to generate images with diverse content and global layout. PetsGAN [30] fully uses external and internal priors within a single-scale GAN framework, achieving good image generation results.

Due to the excellent image generation ability of the diffusion models and their absence of mode collapse compared to GANs, we propose using a diffusion model to handle single images with strong global characteristics. Existing diffusion models typically remove attention layers and use fewer or no up-and-down sampling layers to learn the internal distribution of patches within a single image. This results in a lack of global image structure and lower quality in generated images. Inspired by ExSinGAN, we propose **S**emantic fusion and **S**tructure-guided global generation from a **S**ingle image with **Diff**usion models (**S^{3}Diff**). We introduce a semantic fusion block that fuses semantic features with image features, enabling the model to draw inspiration from high-level semantic features so that the generated image has reasonable semantics. Additionally, during sampling, we employ manifold constrained gradient based on image structure to ensure the generation path regresses to the original image manifold, thereby preserving the structural relationships within the image. We conduct extensive experiments on public datasets, thoroughly examining the rationality of each design through both qualitative and quantitative evaluations, and validate the effectiveness of our proposed method by comparing it with baseline models.

Overall, our main contributions are as follows:

- We use a semantic extractor to extract high-level semantic features from images and propose a semantic fusion block to fuse semantic features with image features so that the generated image has reasonable semantics.
- During the sampling, we use a structural feature extractor to extract structural features from images. By calculating structural gradients, we force the generated image's path to regress to the manifold of the original image, thereby preserving the images' structural relationships.
- We conduct extensive experiments on public datasets, providing a detailed investigation of hyperparameter selection and the rationality of key designs. Our method not only achieves conventional image generation but also accomplishes image manipulation tasks by redesigning the sampling procedure.

2 Related work

2.1 Diffusion models in image generation

[31] is the first to apply diffusion concepts to generative modeling, though it did not attract much attention at the time. DDPM [10] advancing [31] demonstrates impressive generation capabilities from high-level details and the diversity of generated examples. Following this, diffusion models emerge as a class of deep generative models and become one of the hottest topics in computer vision. So far, diffusion models have been successfully applied to a wide range of generative modeling tasks, such as noise-only image generation

[11], image super-resolution [32], image restoration [14], image editing [33], and image-to-image translation [13]. Among these, reverse sampling in diffusion models plays a significant role in determining the quality of generated images, and by appropriately modifying the sampling process and using a guidance function, the quality of the generated images can be improved. [11] train a classifier on images of different noise scales as the guidance function and use the gradient of the classifier as the score function to guide the model in generating images of the user-specified classes during sampling. However, such methods lack generalization, requiring the classifier to be retrained whenever the conditions change. To address this, researchers employ unconditional score functions combined with projection-based measurement constraints to solve various inverse problems in an unsupervised manner [34, 35, 36, 37]. [38] introduces an additional correction term inspired by the manifold constraint, using the Manifold Constrained Gradient (MCG) to enforce the sample path to regress onto a plane tangent to the original data manifold. Inspired by MCG, we incorporate the manifold constrained gradient based on image structure during the sampling process to preserve the global structure of the image.

2.2 Single-image generative models

Using a single image to train a generative model enables the model to learn the internal distribution of patches within a single image and generalize novel patterns from the training images, resulting in realistic and diverse images. SinGAN [1] is the first to use the multi-scale GAN to train a purely deep generation model (i.e., maps noise to image samples) on a single natural image. Subsequently, the multi-scale GAN structure proposed by SinGAN gradually became a paradigm for single-image generative models [2, 3, 5, 28, 4, 6, 39]. Multi-scale GAN employs multiple downsampled versions of the original image, starting from the coarsest scale, and sequentially trains GANs at different scales for image generation. However, this structure suffers from error accumulation, which reduces the quality of generated images [7].

One-shot GAN [29] proposes a single-scale GAN that combines the two-branch discriminator and diversity regularization technique, which can generate new images with reasonable composition and visual realism. PetsGAN [30], based on the concept of single-scale GAN, decomposes the image generator into a small-scale GAN serving as an external prior network and an image super-resolution network acting as an internal prior network. This approach integrates external and internal prior information, preserving the semantic and structural relationships in the generated images. However, GANs, which employ an adversarial training strategy, are prone to mode collapse. Although GPNN [12] replaces GAN with a simple (up-graded) Patch Nearest-Neighbor (PNN) module and achieves higher quality outputs while reducing generation time by orders of magnitude, it is short of generalization and generates images close to the training images [8].

Recently, diffusion models have demonstrated significant potential in single-image generation tasks. SinDDM [9] and SinODE [40] leverage the concept of multi-scale GANs by employing multi-scale diffusion models to generate high-quality and diverse images in a coarse-to-fine manner. On the other hand, SinDiffusion [7] and SinFusion [8] use a single-scale diffusion model, generating realistic and diverse images through a multi-step denoising process. While diffusion model-based single-image generation models enhance the quality and diversity of generated images and avoid model collapse, they struggle to handle images with global structures, such as animal images. Our proposed method incorporates semantic fusion and structural guidance, enabling the generation of images that preserve reasonable semantics and global structures.

3 Method

There is a natural image $\boldsymbol{x}_0$, and the purpose of single-image generation is to use a diffusion model p_θ to learn the internal distribution of patches within a single image $\boldsymbol{x}_0$, such that the distribution of images generated by the model from random noise $\boldsymbol{z}$ is similar to that of image $\boldsymbol{x}_0$. Since single-image generation models focus on learning the internal distribution of patches within a single image during training, they emphasize local features while neglecting some global structures that are essential for some images (e.g., animals), resulting in poor quality of the generated images. To address this

issue, we propose semantic fusion and structure-guided global generation from a single image with diffusion models. We employ a commonly used semantic extractor to extract high-level semantic features from the image and use a semantic fusion block to fuse semantic features with image features. During sampling, we guide the generation of the image using the manifold constrained gradient based on image structure, ultimately generating high-quality and diverse images while preserving reasonable semantics and structure. The specific implementation details are as follows.

3.1 Semantic fusion diffusion model

To enhance the model's understanding of image semantics, we incorporate a semantic fusion block into vanilla single-image generation models to form the semantic fusion diffusion model. The semantic fusion diffusion model we propose is shown in Figure 1(b).

Compared to vanilla single-image generation models, we use a semantic extractor E_{Se} to extract high-level semantic features $\boldsymbol{Se}$ from the image. These features are then fused with the image features through a semantic fusion block, enabling the model to learn more image semantic information. During the forward process, the noisy image

$$\boldsymbol{x}_t \sim q_{t|0}(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\gamma_t}\boldsymbol{x}_0, (1-\gamma_t)\boldsymbol{I}) \quad (1)$$

at time $t \in \{0, 1, \ldots, T\}$ of image $\boldsymbol{x}_0$ is input into the diffusion model p_θ along with the semantic feature $\boldsymbol{Se}$=$\mathrm{E}_{Se}(\boldsymbol{x}_0)$ of image $\boldsymbol{x}_0$, where $q_{t|0}(\cdot\,|\,\cdot)$ is the perturbation kernel, γ_t is the noise schedule at time t, and $\boldsymbol{I}$ is the identity matrix. The loss function for model training is

$$L = \mathbb{E}\|\boldsymbol{\psi} - p_\theta(\boldsymbol{x}_t, t, \mathrm{E}_{Se}(\boldsymbol{x}_0))\|^2, \quad (2)$$

where $p_\theta(\cdot)$ is the output of the diffusion model, $\boldsymbol{\psi}$ is the target for training the model, and $\mathbb{E}$ is the expectation. Typically, $\boldsymbol{\psi}$ can be noise sampled from a standard Gaussian distribution or the original image $\boldsymbol{x}_0$.

3.2 Semantic extractor

An image semantic extractor is required to introduce semantic features of images into vanilla single-image generation models. The representational capability of the semantic extractor directly impacts the correctness and completeness of the extracted image semantics. Benefiting from the development of large-scale image datasets, many pretrained visual models can effectively extract semantic features from images. Among these, the most widely used is Contrastive Language-Image Pretraining (CLIP) [16]. CLIP is a pretrained model using contrastive learning on large-scale language-image pairs, with its Image Encoder specifically designed to extract high-level semantic features from images. Considering the constraints of storage capacity and the need for pixel-wise semantic features, we adopt the pretrained CLIP "RN50" model as the semantic extractor. Since "RN 50" has a total of 4 layers, each extracting different semantic features, with deeper layers extracting more abstract features, we conduct ablation experiments (see Sec. 4.5) to select the most suitable semantic features and experimentally find that the semantic features from the second layer are optimal. We represent the semantic extractor that outputs the optimal semantic features as E_{Se} and the semantic features extracted from image $\boldsymbol{x}_0$ by E_{Se} as $\boldsymbol{Se}$.

3.3 Semantic fusion block

After obtaining the semantic feature $\boldsymbol{Se}$, we propose a semantic fusion block to fuse semantic features with image features. The architecture of the semantic fusion block is shown in Figure 1(c). The image feature $\boldsymbol{f}$ is first passed to the mapping layer to obtain

$$f' = \mathrm{Conv}(\mathrm{GN}(\boldsymbol{f})), \quad (3)$$

where GN represents the group normalization module, and Conv represents the convolution layer. The output feature f' from the mapping layer is then fed into the linear self-attention module to produce the query

$$\boldsymbol{Q} = \mathrm{LN}\left(\mathrm{FF}\left(\mathrm{LN}\left(\mathrm{LSA}\left(\mathrm{LN}\left(f'\right)\right)\right)\right)\right), \quad (4)$$

where LN represents the layer normalization module, LSA represents the linear self-attention module, and FF represents the feedforward module. On the other hand, the semantic features $\boldsymbol{Se}$ of the image, after passing through the convolutional layer, are regarded as the key $\boldsymbol{K}$ and value $\boldsymbol{V}$. Finally, the output

$$f'' = \mathrm{FF}(\mathrm{LN}(\mathrm{LCA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}))) \quad (5)$$

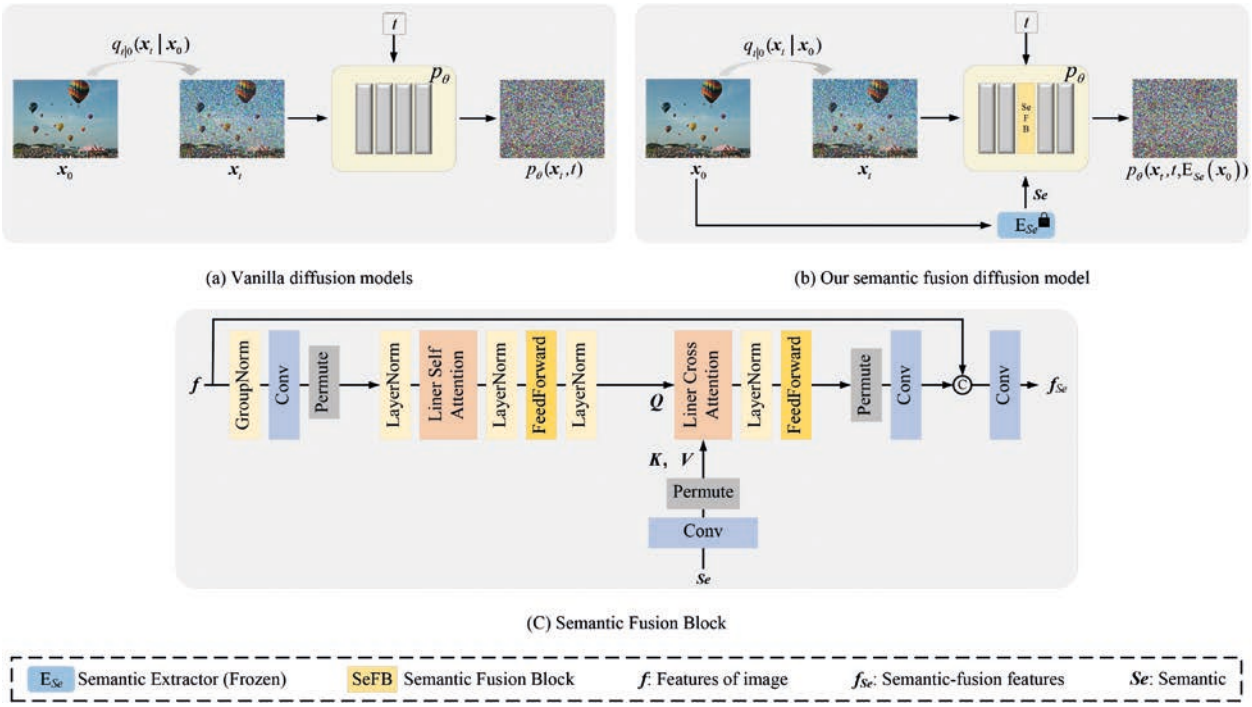


Figure 1. Illustration of (a) Vanilla diffusion models. (b) Our proposed semantic fusion diffusion model. (c) The network structure of SeFB. Vanilla diffusion models do not consider semantic information during training, resulting in unreasonable semantics in the generation of global images. On the contrary, our proposed semantic fusion diffusion model fuses high-level semantic information of images as conditions during training, enabling the model to learn more semantic information from images.

from the linear cross-attention module is concatenated with the image feature $\boldsymbol{f}$ to form the final output

$$\boldsymbol{f}_{Se} = \mathrm{Conv}\left(\mathrm{Concat}\left(\boldsymbol{f}, \mathrm{Conv}\left(f''\right)\right)\right), \quad (6)$$

where LCA represents the linear cross-attention module, and Concat represents the concatenate operation.

Using semantic fusion blocks, the model enhances its understanding of image semantics during training, thereby improving the quality of generated images. Intuitively, the position of the semantic fusion block affects the quality of the generated image. To investigate this, we conduct ablation experiments by inserting the semantic fusion block after the j-th block ($j = 1, ..., N$, where N is the total number of blocks) of the base model (see Sec. 4.5) and find that $j = 8$ is the most reasonable.

3.4 Structure-guided reverse sampling

After the model trained, we generate new images from $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ using the reverse sampling formula

$$\boldsymbol{x}_{t-1} \sim p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \Sigma_\theta(\boldsymbol{x}_t, t)\boldsymbol{I}), \quad (7)$$

where $\mu_\theta(\boldsymbol{x}_t, t)$ and $\Sigma_\theta(\boldsymbol{x}_t, t)$ are the predicted mean and variance from the model, respectively.

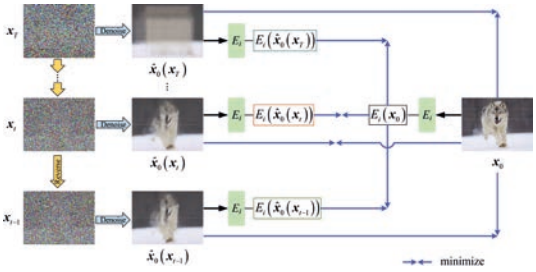


Figure 2. Structure-guided Reverse Sampling. Given the input image $\boldsymbol{x}_0$, we guide the reverse diffusion process to use the manifold constrained gradient (MCG) of structural features in the first t_s moments.

Although we have introduced high-level semantic features of images in vanilla single-image generation models, some generated images still lack global structural relationships. As shown in Figure 2, inspired by the recent manifold constrained gradient (MCG) for inverse problems [38], we introduce the MCG of structural features in the first t_s moments of reverse sampling, guiding the image to retain basic structural relationships by minimizing the cost function

$$\ell_{cost}(\boldsymbol{x}, \boldsymbol{x}_0) = \lambda_{cost} \|E_i(\boldsymbol{x}) - E_i(\boldsymbol{x}_0)\|_2 \quad (8)$$

for the sample $\boldsymbol{x}$, where E_i is the image structure feature extractor and λ_{cost} is a hyperparameter. In experiments, we find that using E_i alone often results in artifacts (see Figure 3). To address this issue, we guide the output to follow the overall structure of the original image with a weak MSE loss between the images. Therefore, our cost function is modified to

$$\ell_{cost}(\boldsymbol{x}, \boldsymbol{x}_0) = \lambda_{cost} \|E_i(\boldsymbol{x}) - E_i(\boldsymbol{x}_0)\|_2 + \lambda_{mse} \|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \quad (9)$$

for the sample $\boldsymbol{x}$, where λ_{mse} is a hyperparameter.

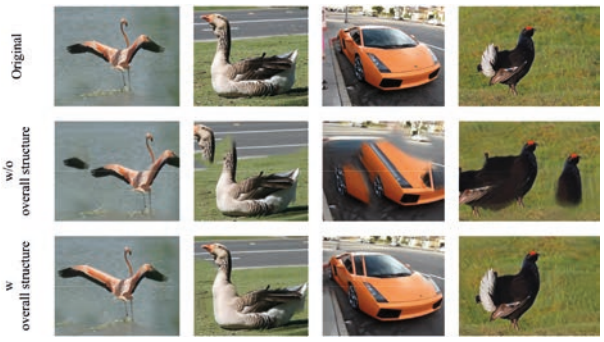


Figure 3. Random samples on Imagenet50. "w/o overall structure" means that the overall structure of the original image is not used, resulting in artifacts in the generated image. "w overall structure" means that the overall structure of the original image is used, resulting in higher-quality images.

Thus, the reverse sampling formula after introducing the MCG of structural features is

$$\begin{aligned} \boldsymbol{x}'_{t-1} &\sim \mathcal{N}(\boldsymbol{x}'_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \Sigma_\theta(\boldsymbol{x}_t, t)\boldsymbol{I}) \\ \boldsymbol{x}_{t-1} &= \boldsymbol{x}'_{t-1} - \nabla_{\boldsymbol{x}_t} \ell_{cost}(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t), \boldsymbol{x}_0) \end{aligned}. \quad (10)$$

When $\boldsymbol{\psi}$ is sampled from the standard Gaussian distribution, $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$ refers to the estimated clean image from the sample $\boldsymbol{x}_t$ using Tweedie's formula

$$\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t) = \frac{1}{\sqrt{\gamma_t}} \left(\boldsymbol{x}_t - \sqrt{1-\gamma_t} p_\theta(\boldsymbol{x}_t, t, \mathrm{E}_{Se}(\boldsymbol{x}_0))\right); \quad (11)$$

When $\boldsymbol{\psi}$ is the original image $\boldsymbol{x}_0$, $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t) = p_\theta(\boldsymbol{x}_t, t, \mathrm{E}_{Se}(\boldsymbol{x}_0))$.

4 Experiments

We conduct extensive experiments on public datasets to explore the rationality of hyperparameter selection and key designs and perform qualitative and quantitative comparisons with baseline methods to validate the effectiveness of our proposed approach.

4.1 Setup

Datasets

For a fair comparison, we use datasets commonly employed in the research field, which mainly include the following:

- Places50. Places50 [1] consists of 50 natural landscape images from different classes and was selected from the Places365 [17] dataset by SinGAN.
- SIGD16. SIGD16 [12] comprises 16 images with distinct structural differences, offering many more conceivable versions per image compared to Places images, and was used by GPNN to evaluate the quality of randomly generated images.
- Lsun50. Lsun50 [3] was created by CosSinGAN taking five random samples from each of the ten classes of the LSUN dataset [41], resulting in 50 images that exhibit a more global layout or structure.
- Imagenet50. Imagenet50 [4] consists of 50 semantic images and was selected by ExSinGAN from the validation set of Imagenet [42] for testing the generation of semantic images.

Evaluation metrics

The quality of generated images is usually evaluated from both qualitative and quantitative perspectives. The quantitative evaluation metrics we use are as follows:

- Single Image Frechet Inception Distance (SIFID). SIFID [1] evaluates the quality of a single generated image by measuring the difference between the feature vectors of the generated image $\boldsymbol{x}_g$ and real image $\boldsymbol{x}_r$. When calculating SIFID, use deep features at the output of the convolutional layer just before the second pooling layer in the Inception Network [44]. The calculation formula is

$$\begin{aligned}\mathrm{SIFID}(\boldsymbol{x}_g,\boldsymbol{x}_r) =& \left\|\mu_{\boldsymbol{x}_g},\mu_{\boldsymbol{x}_r}\right\|^2 \\ &+tr\left(\boldsymbol{\Sigma}_{\boldsymbol{x}_g}+\boldsymbol{\Sigma}_{\boldsymbol{x}_r}-2\cdot\left(\boldsymbol{\Sigma}_{\boldsymbol{x}_g}\cdot\boldsymbol{\Sigma}_{\boldsymbol{x}_r}\right)^{\frac{1}{2}}\right),\end{aligned} \tag{12}$$

  where $\mu_{\boldsymbol{x}_g}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}_g}$ represent the mean and covariance of the feature vectors of the generated images, and $\mu_{\boldsymbol{x}_r}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}_r}$ represent the mean and covariance of real image feature vectors. A lower SIFID value indicates higher image quality.
- Learned Perceptual Image Patch Similarity (LPIPS). LPIPS [15] primarily measures the perceptual similarity between generated images and is used to evaluate the diversity of generated images. When calculating LPIPS, the generated images $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are input into a pre-trained classification model (e.g., VGG [43]) to obtain the output features $\widehat{\boldsymbol{y}}_1^l,\widehat{\boldsymbol{y}}_2^l\in\mathbb{R}^{H_l\times W_l\times C_l}$ from the l-th layer. Then, the differences between the features are computed, with the calculation formula being

$$\mathrm{LPIPS}\left(\boldsymbol{y}_1,\boldsymbol{y}_2\right)=\sum_l\frac{1}{H_lW_l}\sum_{h,w}\left\|\boldsymbol{w}_l\odot\left(\widehat{\boldsymbol{y}}_{1hw}^l-\widehat{\boldsymbol{y}}_{2hw}^l\right)\right\|_2^2, \tag{13}$$

  where l is the layer index of the classification model, $\widehat{\boldsymbol{y}}_{1hw}^l$ and $\widehat{\boldsymbol{y}}_{2hw}^l$ are the values of $\widehat{\boldsymbol{y}}_1^l$ and $\widehat{\boldsymbol{y}}_2^l$ at the coordinate point (h,w), $\boldsymbol{w}_l$ is the trainable weight of the l-th layer, $\odot$ represents the dot product, and H_l and W_l denotes the height and width of the feature vector from the l-th layer. For a set of generated images $\boldsymbol{y}^N=\{\boldsymbol{y}_i\}_{i=1}^N$, the calculation formula is

$$\mathrm{LPIPS}\left(\boldsymbol{y}^N\right)=\frac{2}{N(N-1)}\sum_{i=1}^N\sum_{j=i+1}^N\sum_l\frac{1}{H_lW_l}\sum_{h,w}\left\|\boldsymbol{w}_l\odot\left(\widehat{\boldsymbol{y}}_{1hw}^l-\widehat{\boldsymbol{y}}_{2hw}^l\right)\right\|_2^2. \tag{14}$$

  The higher the LPIPS value, the richer the diversity of the generated images in this set.

Implementation details

We use SinFusion [8] as the experimental base model, and the structure feature extractor employs a low-pass filter. The whole framework is implemented by Pytorch, and the experiments are performed on a single NVIDIA A100. When training the Imagenet50 dataset, the longest side of each image is scaled to 256, with the shorter side scaled proportionally, while the resolution of images in other datasets remains unchanged. Other training parameters are consistent with SinFusion.

Reproductions

All baselines are reproduced based on the public code, as shown in Table 1. SinDiffusion, after training on the Imagenet50 dataset, generates images with the longest side resized to 250, consistent with the training process. SinFusion trains the Imagenet50 dataset with each image's longest side resized to 256.

Table 1. The used codes.

Method	URL
SinGAN [1]	https://github.com/tamarott/SinGAN
One-shot GAN [29]	https://github.com/boschresearch/one-shot-synthesis
ExSinGAN [4]	https://github.com/zhangzc21/ExSinGAN
PetsGAN [30]	https://github.com/zhangzc21/PetsGAN
SinDiffusion [7]	https://github.com/WeilunWang/SinDiffusion
SinDDM [9]	https://github.com/fallenshock/SinDDM
SinFusion [8]	https://github.com/yanivnik/sinfusion-code

4.2 Qualitative evaluation

We train S^{3}Diff on Places50 [1], SIGD16 [12], Lsun50 [3], and Imagenet50 [4], then generate images using structure-guided reverse sampling, as shown in Figure 4.

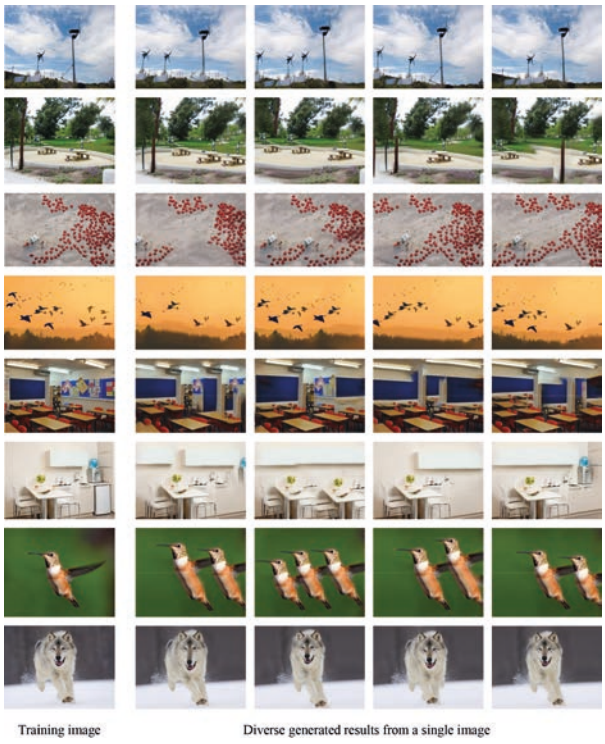


Figure 4. Random samples from single images. It can be seen that our method is capable of handling both small target images and images with global features, generating high-quality and diverse images.

From Figure 4, it can be seen that our method can not only generate small-target images but also generate global images with reasonable semantics and structural relationships while maintaining diversity. To demonstrate the superiority of S^{3}Diff, we compare our S^{3}Diff with several challenging methods, i.e., SinGAN [1], One-shot GAN [29], ExSinGAN [4], PetsGAN [30], SinDiffusion [7], SinDDM [9], SinFusion [8], as shown in Figure 5, where red boxes indicate lower quality images.

As seen in Figure 5, the GAN-based baseline models generate images with unreasonable semantic and structural relationships, where images generated by SinGAN have artifacts; One-shot GAN fails to preserve structural relationships in images generated on Imagenet50, and the images generated on Places50, SIGD16, and Lsun50 are close to the original image and lack diversity (see Table 2 for details); ExSinGAN and PetsGAN generate images with artifacts on Lsun50. The visual quality of images generated by diffusion-based baseline models is also inconsistent. Many images generated by SinDiffusion and SinDDM have artifacts. Additionally, images generated by SinFusion also exhibit artifacts and lose the main subject when generating images on Imagenet50. In contrast, our S^{3}Diff can generate high-quality and diverse images while preserving reasonable semantic and structural relationships.

4.3 Quantitative evaluation

In addition to qualitative comparisons, we also quantitatively compare our S^{3}Diff with several challenging methods, as shown in Table 2. Each image generates one image for SIFID calculation and ten for LPIPS calculation.

As shown in Table 2, our S^{3}Diff has a lower SIFID and higher LPIPS compared to GAN models such as SinGAN, One-shot GAN, and PetsGAN. This indicates that S^{3}Diff generates high-quality and diverse images, while the latter models exhibit a certain degree of mode collapse, particularly One-shot GAN, resulting in a lack of diversity in the generated images. Compared with S^{3}Diff, although ExSinGAN has a significantly lower SIFID score on Places50 and Lsun50, its generated images have poorer visual quality (see Figure 5), and its LPIPS score is also notably lower than that of S^{3}Diff. Compared with diffusion-based baseline models, S^{3}Diff not only has significantly lower SIFID but

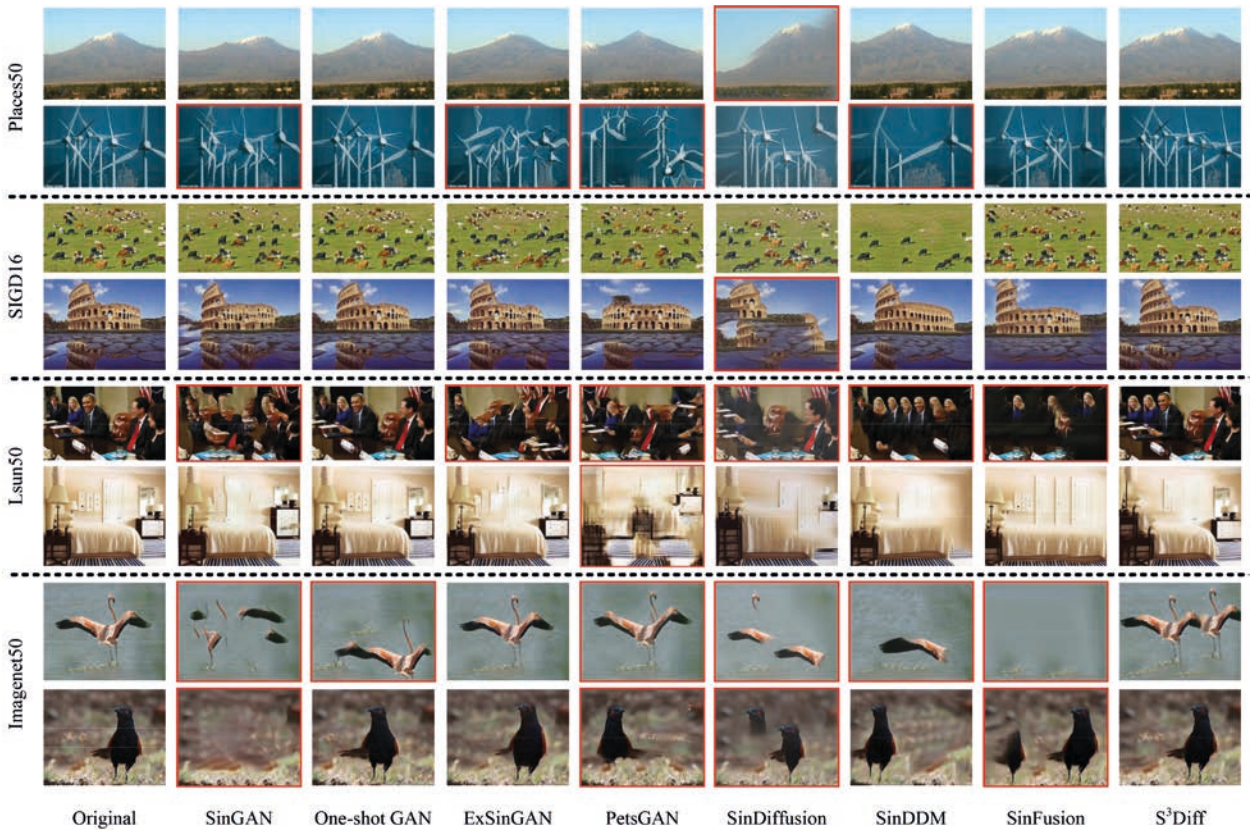


Figure 5. Qualitative comparison on different dataset. We compare our S^3Diff with several challenging methods, i.e., SinGAN [1], One-shot GAN [29], ExSinGAN [4], PetsGAN [30], SinDiffusion [7], SinDDM [9], SinFusion [8]. Poor-quality images are highlighted with red boxes. It can be seen that the images generated by the baseline model have unreasonable semantic and structural relationships, while our S^3Diff can generate high-quality and diverse images while preserving reasonable semantic and structural relationships.

Table 2. Quantitative comparison on different dataset. ↑ indicates the higher the better, while ↓ indicates the lower the better.

Method	Places50		SIGD16		Lsun50		Imagenet50	
	SIFID↓	LPIPS↑	SIFID↓	LPIPS↑	SIFID↓	LPIPS↑	SIFID↓	LPIPS↑
SinGAN [1]	0.14	0.262	0.18	0.322	0.17	0.367	0.69	0.346
One-shot GAN [29]	0.21	0.026	0.21	0.065	0.18	0.033	0.34	0.201
ExSinGAN [4]	0.08	0.229	0.09	0.227	0.07	0.268	0.43	0.222
PetsGAN [30]	0.16	0.293	0.10	0.312	0.14	0.395	0.55	0.359
SinDiffusion [7]	0.49	0.429	0.44	0.445	0.51	0.507	0.91	0.486
SinDDM [9]	0.52	0.335	1.08	0.365	1.38	0.427	1.11	0.365
SinFusion [8]	0.61	0.363	0.37	0.365	0.74	0.444	1.21	0.430
S^3Diff (Ours)	0.19	0.327	0.11	0.368	0.12	0.423	0.43	0.398

also competitive LPIPS, with visually superior generated images (see Figure 5), demonstrating that our S^{3}Diff can generate high-quality and diverse images while preserving reasonable semantic and structural relationships.

4.4 Image manipulation

In addition to conventional image generation, our method can also achieve many image manipulation tasks, such as editing, style transfer, and paint-to-image translation, as shown in Figure 6.

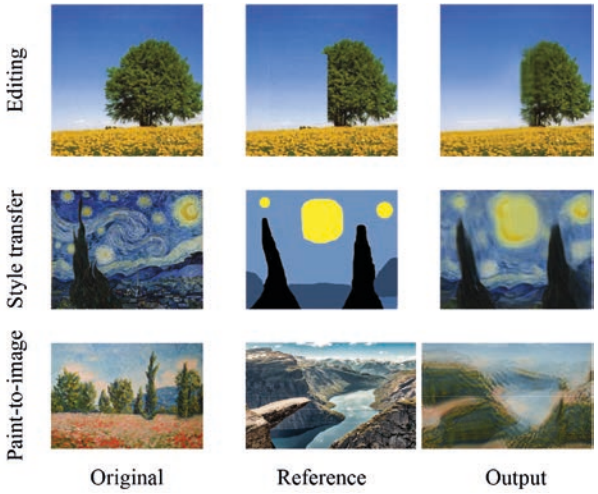


Figure 6. Application of image manipulation task.

After training our proposed model on the original image, we perform image manipulation based on the reference image using our proposed structure-guided reverse sampling. Specifically, the training process remains unchanged, and during sampling, $t_s = T$, with the sampling formula being

$$\begin{aligned} \boldsymbol{x'}_{t-1} &\sim \mathcal{N}\left(\boldsymbol{x'}_{t-1}; \mu_\theta\left(\boldsymbol{x}_t, t\right), \Sigma_\theta\left(\boldsymbol{x}_t, t\right) \boldsymbol{I}\right) \\ \boldsymbol{x}_{t-1} &= \boldsymbol{x'}_{t-1} - \nabla_{\boldsymbol{x}_t} \ell_{cost}\left(\hat{\boldsymbol{x}}_0\left(\boldsymbol{x}_t\right), \boldsymbol{x}^*\right) \end{aligned}, \quad (15)$$

where $\boldsymbol{x}^*$ is the reference image, and * represents editing, style transfer, and paint-to-image translation.

4.5 Ablation study

We select reasonable hyperparameters through ablation experiments and explore the effectiveness of the proposed key designs.

Selection of the index j

To determine the specific layer index j for inserting the semantic fusion block into SinFusion, we selected one image each from the SIGD16 and Imagenet50 datasets as training images and conducted experiments by sequentially inserting the semantic fusion block after each block in SinFusion. During sampling, t_s is set to 1, and λ_{cost} and λ_{mse} are set to 0.05. The semantic features used the output from the second layer of CLIP "RN 50". The qualitative results are shown in Figure 7.

As shown in Figure 7, except for $j \in \{7,8,9\}$, inserting semantic fusion blocks into other layers results in lower image quality. For example, the images generated from $\boldsymbol{I}_2$ exhibit varying degrees of artifacts, while those from $\boldsymbol{I}_1$ either lack the main object (large colored balloon) or contain artifacts. To further select an appropriate j, we conducted a quantitative comparison on $j \in \{7,8,9\}$, generating 50 images for each to compute all metrics, with the results presented in Table 3.

Table 3. Quantitative evaluation results for different index j. We evaluate the images generated by $j \in \{7,8,9\}$ using quantitative evaluation metrics.

Index	$\boldsymbol{I}_1$		$\boldsymbol{I}_2$	
	SIFID↓	LPIPS↑	SIFID↓	LPIPS↑
$j=7$	0.43	0.371	0.27	0.142
$j=8$	0.20	0.407	0.17	0.145
$j=9$	0.40	0.367	0.26	0.153

As shown in Table 3, image $\boldsymbol{I}_1$ achieves the optimal metric value when $j=8$, image $\boldsymbol{I}_2$ reaches the optimal SIFID value when $j=8$, while LPIPS is slightly worse than $j=9$. Combining Figure 7 and Table 3, we ultimately select $j=8$ as the index for the semantic fusion block.

The influence of hyperparameters λ_{cost} and λ_{mse}

We set different values of λ_{cost} and λ_{mse} on Imagenet50, and t_s is set to 1. The experimental results are shown in Table 4.

From Table 4, it can be observed that as $\lambda_{mse} = 0$ increases with λ_{cost}, the SIFID of the generated images fluctuates increases while the LPIPS gradually decreases, indicating a decline in image quality and a lack of diversity. When $\lambda_{cost} = 0.05$, as λ_{mse} increases, the SIFID of the generated images

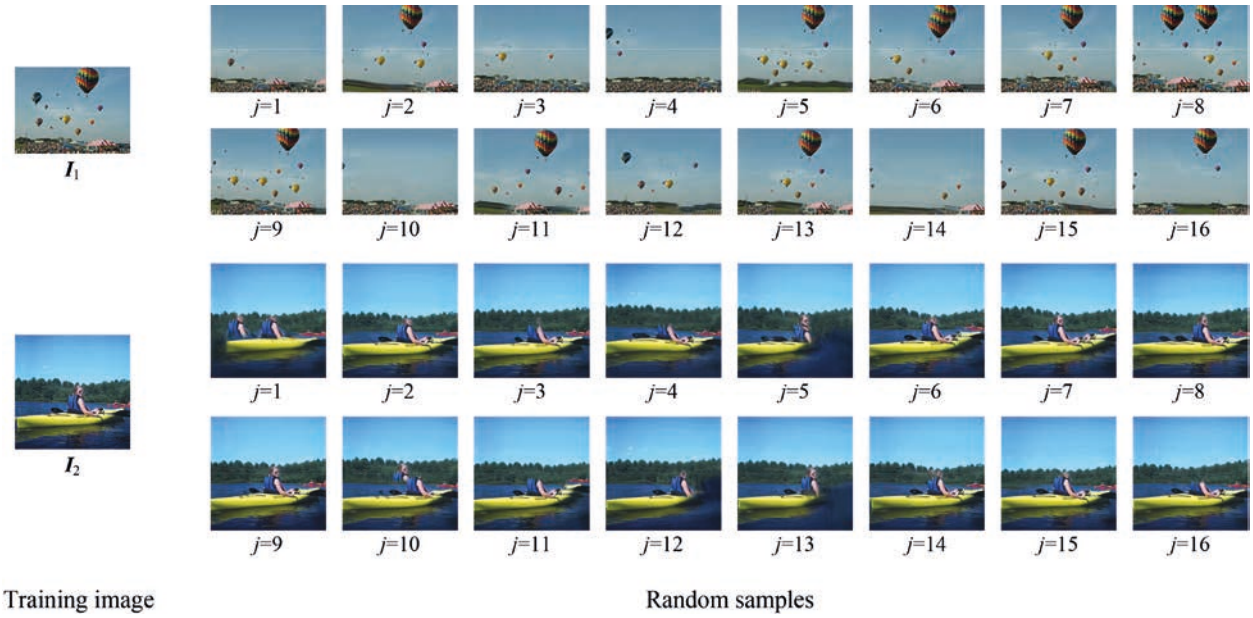


Figure 7. The images generated by different index j. We insert the semantic fusion block into different layer index j of SinFusion and use images $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ to train the model.

gradually rises, and the LPIPS gradually decreases, suggesting that both the quality and diversity of the generated images are decreasing.

Table 4. Quantitative evaluation results of different λ_{cost} and λ_{mse} on Imagenet50.

$\lambda_{cost}, \lambda_{mse}$	Imagenet50	
	SIFID↓	LPIPS↑
$\lambda_{cost}=0, \lambda_{mse}=0$	0.65	0.417
$\lambda_{cost}=0.05, \lambda_{mse}=0$	0.61	0.403
$\lambda_{cost}=0.1, \lambda_{mse}=0$	0.66	0.377
$\lambda_{cost}=0.2, \lambda_{mse}=0$	0.60	0.332
$\lambda_{cost}=0.3, \lambda_{mse}=0$	0.69	0.319
$\lambda_{cost}=0.05, \lambda_{mse}=0.05$	0.43	0.398
$\lambda_{cost}=0.05, \lambda_{mse}=0.1$	0.45	0.318
$\lambda_{cost}=0.05, \lambda_{mse}=0.2$	0.47	0.261
$\lambda_{cost}=0.05, \lambda_{mse}=0.3$	0.48	0.243

The selection of hyperparameter t_s

Since t_s directly determines our constraint on the global structure of the generated image, we set different t_s and obtain the experimental results shown in Figure 8.

As shown in Figure 8, as t_s increases, SIFID and LPIPS generally show a downward trend. It is because the generated images become increasingly similar to the original images, resulting in a lack of diversity. Due to the significant decrease in LPIPS, however, to ensure the generated images maintain rich diversity, we ultimately chose $t_s = 1$ as the timestep for structural guidance.

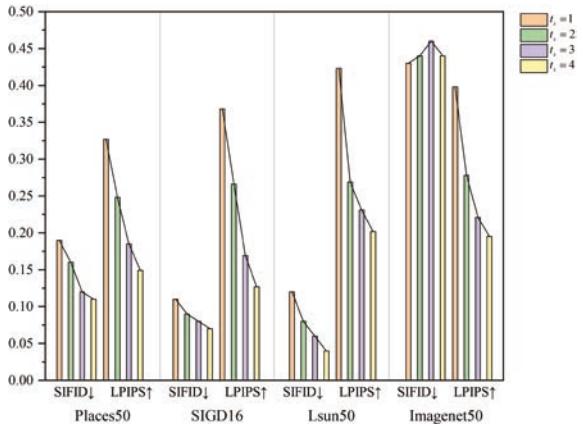


Figure 8. Quantitative evaluation results for different t_s. We set different t_s to generate images on Places50, SIGD16, Lsun50, and Imagenet50 to compute SIFID and LPIPS. It can be seen that LPIPS decreased more than SIFID.

The semantics of different layers

The CLIP "RN 50" model consists of 4 layers, with the abstraction of semantic information extracted from images increasing as the number of layers increases. We utilize the image features ex-

tracted by each layer as semantic features input to the semantic fusion block. The quantitative evaluation results generated on Lsun50 and Imagenet50 are shown in Table 5.

Table 5. Quantitative evaluation results of different layers. We use the output of each layer from CLIP "RN 50" as image semantic features.

Method	Lsun50		Imagenet50	
	SIFID↓	LPIPS↑	SIFID↓	LPIPS↑
S^3Diff-Layer1	0.17	0.422	0.49	0.384
S^3Diff-Layer2	0.12	0.423	0.43	0.398
S^3Diff-Layer3	0.14	0.420	0.47	0.379
S^3Diff-Layer4	0.20	0.432	0.44	0.398

As shown in Table 5, when using the output of the second layer as semantic features to generate images, SIFID is the lowest, while using the output of the fourth layer as semantic features results in the highest LPIPS. When comparing S^3Diff-Layer2 and S^3Diff-Layer4 on Lsun50, although SIFID increases by 67%, LPIPS only increases by 2%. Therefore, we ultimately choose the output of the second layer as the semantic feature.

The influence of key designs

In this paper, we primarily propose two key designs: first, inserting a semantic fusion block into the base model, and second, introducing structure-guided reverse sampling. For the effectiveness of the above two designs, the experimental results are shown in Table 6.

Table 6. Evaluation results of different designs. "w/o SeFB" means not using semantic fusion blocks, and "w/o SRS" means not using structure-guided reverse sampling.

Method	Places50		SIGD16		Lsun50		Imagenet50	
	SIFID↓	LPIPS↑	SIFID↓	LPIPS↑	SIFID↓	LPIPS↑	SIFID↓	LPIPS↑
w/o SeFB	0.71	0.343	0.43	0.363	0.35	0.432	0.76	0.388
w/o SRS	0.39	0.347	0.19	0.382	0.43	0.432	0.65	0.417
S^3Diff (Ours)	0.19	0.327	0.11	0.368	0.12	0.423	0.43	0.398

As indicated in Table 6, the SIFID of the generated images significantly increases without semantic fusion, while the LPIPS changes slightly, suggesting that semantic fusion can improve the quality of the generated images. Without structure-guided reverse sampling, both the SIFID and LPIPS of the generated images increase, but the generated images lack reasonable structural relationships (see Figure 3). In summary, both key designs we propose are essential.

4.6 The computation cost

We select one image from each of Places50, SIGD16, Lsun50, and Imagenet50 (as shown in Figure 9) to evaluate the computation cost. We compare the computation cost of SinFusion [8] with our S^3Diff using four metrics: trainable parameters, training time, GPU memory, and sampling time, as shown in Table 7, where the batch size during sampling is set to 50.



Figure 9. Selected images.

Table 7. Computation cost comparison.

Method	Trainable parameters (M)	Training time (s)	GPU memory (MB)	Sampling time (s)
$I_0 : 188 \times 250$				
SinFusion [8]	4.6	2346	2768	41
S^3Diff (Ours)	4.8	3087	3590	51
$I_1 : 188 \times 250$				
SinFusion [8]	4.6	2333	2768	40
S^3Diff (Ours)	4.8	3023	3590	51
$I_2 : 251 \times 168$				
SinFusion [8]	4.6	2554	2360	37
S^3Diff (Ours)	4.8	2977	3040	46
$I_3 : 254 \times 256$				
SinFusion [8]	4.6	4089	4218	56
S^3Diff (Ours)	4.8	5102	5500	71

As indicated in Table 7, due to the addition of a semantic fusion block in SinFusion, there is a slight increase in trainable parameters, which also leads to a slight increase in training time and GPU memory. Additionally, the inclusion of structure guidance during sampling increased the sampling time. Although the computation cost has increased, S^3Diff significantly improves the quality of generated images while preserving reasonable semantic and structural relationships, and we believe this is worthwhile.

Conclusion

In this paper, we address the challenge of global image generation in single-image generation models by proposing **S^3Diff**—**S**emantic fusion and **S**tructure-guided global generation from a **S**ingle

image with **Diff**usion models. To enable the model to generate images with reasonable semantics, we employ a semantic extractor to extract the semantics of the image and utilize our proposed semantic fusion block to fuse high-level semantic features with image features. At the same time, to preserve the structural relationships of the image, we introduce a structure-guided reverse sampling during the sampling process, incorporating manifold constrained gradient based on image structure to enforce the sampling path to regress to the manifold of the original image. We conduct extensive experiments on public datasets, thoroughly investigating hyperparameter selection and the rationality of key designs. Beyond conventional image generation, we achieve image manipulation tasks by redesigning the sampling procedure. By comparing qualitatively and quantitatively with the baseline model, our method demonstrates state-of-the-art visual quality and competitive quantitative metrics, validating its ability to preserve reasonable semantic and structural relationships while generating high-quality and diverse images.

Acknowledgments

The research leading to these results received funding from the National Natural Science Foundation of China (No.62006240), Natural Science Basic Research Program of Shaanxi Province (No.2024JC-YBQN-0664), and China Postdoctoral Science Foundation (No.2024M754276).

References

[1] Shaham, Tamar Rott and Dekel, Tali and Michaeli, Tomer, SinGAN: Learning a Generative Model From a Single Natural Image, 2019, 4569-4579.

[2] Gur, Shir and Benaim, Sagie and Wolf, Lior, Hierarchical Patch VAE-GAN: Generating Diverse Videos from a Single Sample, Advances in Neural Information Processing Systems, 33, 2020, 16761-16772.

[3] Hinz, Tobias and Fisher, Matthew and Wang, Oliver and Wermter, Stefan, Improved Techniques for Training Single-Image GANs, IEEE Winter Conference on Applications of Computer Vision (WACV), 2021, 1300-1309.

[4] Zhang, ZiCheng and Han, CongYing and Guo, TianDe, Exsingan: Learning an explainable generative model from a single image, British machine vision conference (BMVC), 2021.

[5] Chen, Jinshu and Xu, Qihui and Kang, Qi and Zhou, MengChu, MOGAN: Morphologic-Structure-Aware Generative Learning From a Single Image, IEEE Transactions on Systems, Man, and Cybernetics: Systems, 54, 2024, 2021-2033.

[6] Chen, Xi and Zhao, Hongdong and Yang, Dongxu and Li, Yueyuan and Kang, Qing and Lu, Haiyan, SA-SinGAN: self-attention for single-image generation adversarial networks, Mach. Vision Appl., 32, 2021.

[7] Wang, Weilun and Bao, Jianmin and Zhou, Wengang and Chen, Dongdong and Chen, Dong and Yuan, Lu and Li, Houqiang, SinDiffusion: Learning a Diffusion Model From a Single Natural Image, IEEE Transactions on Pattern Analysis and Machine Intelligence, 47, 2025, 3412-3423.

[8] Nikankin, Yaniv and Haim, Niv and Irani, Michal, SinFusion: Training Diffusion Models on a Single Image or Video, International Conference on Machine Learning, 202, 2023, 26199–26214.

[9] Kulikov, Vladimir and Yadin, Shahar and Kleiner, Matan and Michaeli, Tomer, Sinddm: A single image denoising diffusion model, International Conference on Machine Learning, 202, 2023, 17920-17930.

[10] Ho, Jonathan and Jain, Ajay and Abbeel, Pieter, Denoising Diffusion Probabilistic Models, Advances in Neural Information Processing Systems, 33, 2020, 6840-6851.

[11] Dhariwal, Prafulla and Nichol, Alexander, Diffusion Models Beat GANs on Image Synthesis, Advances in Neural Information Processing Systems, 34, 2021, 8780-8794.

[12] Granot, Niv and Feinstein, Ben and Shocher, Assaf and Bagon, Shai and Irani, Michal, Drop the GAN: In Defense of Patches Nearest Neighbors as Single Image Generative Models, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 13450-13459.

[13] Li, Bo and Xue, Kaitao and Liu, Bin and Lai, Yu-Kun, BBDM: Image-to-Image Translation with Brownian Bridge Diffusion Models, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023, 1952-1961.

[14] Lugmayr, Andreas and Danelljan, Martin and Romero, Andres and Yu, Fisher and Timofte, Radu and Van Gool, Luc, RePaint: Inpainting using Denoising Diffusion Probabilistic Models, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 11451-11461.

[15] Zhang, Richard and Isola, Phillip and Efros, Alexei A. and Shechtman, Eli and Wang, Oliver, The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, 586-595.

[16] Radford, Alec and Kim, Jong Wook and Hallacy, Chris and Ramesh, Aditya and Goh, Gabriel and Agarwal, Sandhini and Sastry, Girish and Askell, Amanda and Mishkin, Pamela and Clark, Jack and Krueger, Gretchen and Sutskever, Ilya, Learning Transferable Visual Models From Natural Language Supervision, International Conference on Machine Learning, 139, 2021, 8748-8763.

[17] Zhou, Bolei and Lapedriza, Agata and Khosla, Aditya and Oliva, Aude and Torralba, Antonio, Places: A 10 Million Image Database for Scene Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, 40, 2018, 1452-1464.

[18] Zhou Wang and Yinglin Yu and Zhang, D., Best neighborhood matching: an information loss restoration technique for block-based image coding systems, IEEE Transactions on Image Processing, 7, 1998, 1056-1061.

[19] Pérez, Patrick and Gangnet, Michel and Blake, Andrew, Poisson image editing, ACM Trans. Graph., 22, 2003, 313–318.

[20] Simakov, Denis and Caspi, Yaron and Shechtman, Eli and Irani, Michal, Summarizing visual data using bidirectional similarity, IEEE Conference on Computer Vision and Pattern Recognition, 2008, 1-8.

[21] Wang, Zhengwei and She, Qi and Ward, Tomás E., Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy, ACM Comput. Surv., 54, 2021, 1-38.

[22] Guillermo Iglesias and Edgar Talavera and Alberto Díaz-Álvarez, A survey on GANs for computer vision: Recent research, analysis and taxonomy, Computer Science Review, 48, 2023, 100553.

[23] Melnik, Andrew and Miasayedzenkau, Maksim and Makaravets, Dzianis and Pirshtuk, Dzianis and Akbulut, Eren and Holzmann, Dennis and Renusch, Tarek and Reichert, Gustav and Ritter, Helge, Face Generation and Editing With StyleGAN: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, 46, 2024, 3557-3576.

[24] Croitoru, Florinel-Alin and Hondru, Vlad and Ionescu, Radu Tudor and Shah, Mubarak, Diffusion Models in Vision: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, 45, 2023, 10850-10869.

[25] Bie, Fengxiang and Yang, Yibo and Zhou, Zhongzhu and Ghanem, Adam and Zhang, Minjia and Yao, Zhewei and Wu, Xiaoxia and Holmes, Connor and Golnari, Pareesa and Clifton, David A. and He, Yuxiong and Tao, Dacheng and Song, Shuaiwen Leon, RenAIssance: A Survey Into AI Text-to-Image Generation in the Era of Large Model, IEEE Transactions on Pattern Analysis and Machine Intelligence, 47, 2025, 2212-2231.

[26] He, Chunming and Shen, Yuqi and Fang, Chengyu and Xiao, Fengyang and Tang, Longxiang and Zhang, Yulun and Zuo, Wangmeng and Guo, Zhenhua and Li, Xiu, Diffusion Models in Low-Level Vision: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, 47, 2025, 4630-4651.

[27] Amirhossein Kazerouni and Ehsan Khodapanah Aghdam and Moein Heidari and Reza Azad and Mohsen Fayyaz and Ilker Hacihaliloglu and Dorit Merhof, Diffusion models in medical imaging: A comprehensive survey, Medical Image Analysis, 88, 2023, 102846.

[28] Eyyup Yildiz and Mehmet Erkan Yuksel and Selcuk Sevgen, A single-image GAN model using self-attention mechanism and DenseNets, Neurocomputing, 596, 2024, 127873.

[29] Sushko, Vadim and Gall, Jürgen and Khoreva, Anna, One-Shot GAN: Learning to Generate Samples from Single Images and Videos, IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021, 2596-2600.

[30] Zhang, Zicheng and Liu, Yinglu and Han, Congying and Shi, Hailin and Guo, Tiande and Zhou, Bowen, Petsgan: Rethinking priors for single image generation, Proceedings of the AAAI Conference on Artificial Intelligence, 36, 2022, 3408–3416.

[31] Sohl-Dickstein, Jascha and Weiss, Eric and Maheswaranathan, Niru and Ganguli, Surya, Deep Unsupervised Learning using Nonequilibrium Thermodynamics, International Conference on Machine Learning, 37, 2015, 2256–2265.

[32] Saharia, Chitwan and Ho, Jonathan and Chan, William and Salimans, Tim and Fleet, David J. and Norouzi, Mohammad, Image Super-Resolution via Iterative Refinement, IEEE Transactions on Pattern Analysis and Machine Intelligence, 45, 2023, 4713-4726.

[33] Chenlin Meng and Yutong He and Yang Song and Jiaming Song and Jiajun Wu and Jun-Yan Zhu and Stefano Ermon, SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations, International Conference on Learning Representations, 2022.

[34] Choi, Jooyoung and Kim, Sungwon and Jeong, Yonghyun and Gwon, Youngjune and Yoon, Sungroh, ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models, IEEE International Conference on Computer Vision (ICCV), 2021, 14347-14356.

[35] Chung, Hyungjin and Sim, Byeongsu and Ye, Jong Chul, Come-Closer-Diffuse-Faster: Accelerating Conditional Diffusion Models for Inverse Problems through Stochastic Contraction, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 12403-12412.

[36] Kadkhodaie, Zahra and Simoncelli, Eero, Stochastic Solutions for Linear Inverse Problems using the Prior Implicit in a Denoiser, Advances in Neural Information Processing Systems, 34, 2021, 13242–13254.

[37] Yang Song and Jascha Sohl-Dickstein and Diederik P Kingma and Abhishek Kumar and Stefano Ermon and Ben Poole, Score-Based Generative Modeling through Stochastic Differential Equations, International Conference on Learning Representations, 2021.

[38] Chung, Hyungjin and Sim, Byeongsu and Ryu, Dohoon and Ye, Jong Chul, Improving Diffusion Models for Inverse Problems using Manifold Constraints, Advances in Neural Information Processing Systems, 35, 2022, 25683–25696.

[39] Elnekave, Ariel and Weiss, Yair, Generating Natural Images with Direct Patch Distributions Matching, European Conference on Computer Vision, 2022, 544–560.

[40] Jian Yue and Yan Gan and Lihua Zhou and Yu Zhao and Shuaifeng Li and Mao Ye, ODE-based generative modeling: Learning from a single natural image, Expert Systems with Applications, 276, 2025, 127185.

[41] Yu, Fisher and Zhang, Yinda and Song, Shuran and Seff, Ari and Xiao, Jianxiong, LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop, arXiv preprint arXiv:1506.03365, 2015.

[42] Deng, Jia and Dong, Wei and Socher, Richard and Li, Li-Jia and Kai Li and Li Fei-Fei, ImageNet: A large-scale hierarchical image database, IEEE Conference on Computer Vision and Pattern Recognition, 2009, 248-255.

[43] Simonyan, K. and Zisserman, A., Very Deep Convolutional Networks for Large-Scale Image Recognition, International Conference on Learning Representations, 2015.

[44] Szegedy, Christian and Wei Liu and Yangqing Jia and Sermanet, Pierre and Reed, Scott and Anguelov, Dragomir and Erhan, Dumitru and Vanhoucke, Vincent and Rabinovich, Andrew, Going deeper with convolutions, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, 1-9.



Xianjie Zhang was born in Mianyang, Sichuan, China, in 1991. He is currently pursuing the Ph.D. degree in computer science and technology with the Xi`an High-Tech research Institute, Xi`an. His research interests include image generation, diffusion model, and computer vision.
https://orcid.org/0009-0004-6049-8506



Yusen Zhang received the B.S. degree in fire control and command control engineering from the Xi`an High-Tech research Institute, Xi`an, China, in 2017, where he is currently working toward the Ph.D. degree in computer science and technology. His research interests include image generation, diffusion model, and computer vision.
https://orcid.org/0009-0004-2268-1510



Yujie He received the B.S. degree in software engineering from Xidian University, Xi`an, China, in 2009, and the Ph.D. degree in intelligent information processing from the Xi`an High-Tech research Institute, Xi`an, China, in 2016. His research interests include infrared small target detection, sparse signal representation, infrared target tracking, generative AI, and cryptography.
https://orcid.org/0000-0002-2299-4945



Min Li received the B.S. and M.S. degrees in computer science and technology from the Xi`an High-Tech research Institute, Xi`an, China, in 1991 and 1994, respectively, and the Ph.D. degree in communication and information from Xi'an Jiaotong University, Xi`an, China, in 1998. She is currently a Professor with the Xi`an High-Tech research Institute. Her research interests include remote sensing, pattern recognition, and generative AI.
https://orcid.org/0000-0002-3009-279X