Learning to Manipulate Artistic Images

Wei Guo*, Yuqi Zhang*, De Ma†, Qian Zheng†

Zhejiang University
{snailforce, yq_zhang, made, qianzheng}@zju.edu.cn

Abstract

Recent advancement in computer vision has significantly lowered the barriers to artistic creation. Exemplar-based image translation methods have attracted much attention due to flexibility and controllability. However, these methods hold assumptions regarding semantics or require semantic information as the input, while accurate semantics is not easy to obtain in artistic images. Besides, these methods suffer from cross-domain artifacts due to training data prior and generate imprecise structure due to feature compression in the spatial domain. In this paper, we propose an arbitrary Style Image Manipulation Network (SIM-Net), which leverages semantic-free information as guidance and a *region transportation* strategy in a self-supervised manner for image generation. Our method balances computational efficiency and high resolution to a certain extent. Moreover, our method facilitates zero-shot style image manipulation. Both qualitative and quantitative experiments demonstrate the superiority of our method over state-of-the-art methods. Code is available at https://github.com/SnailForce/SIM-Net.

Introduction

Art can cultivate sentiment, improve self-cultivation, and inherit culture. The use of artificial intelligence in the process of creating art was significantly accelerated with rapid advances in machine learning. Artificial Intelligence (AI) has endowed art with more possibilities in various artistic fields, such as calligraphy, imagery, design, and others, thereby bridging the gap between people and art. Particularly, image manipulation has unique creativity and strong interactivity, thereby further lowering the barrier to the artistic creation.

Image manipulation essentially falls into the category of image-to-image translation, which has demonstrated success across a wide range of applications (Isola et al. 2017; Liu, Breuel, and Kautz 2017). Different from other conditional image-to-image translation methods, exemplar-based image translation enables more flexible user control and better generation quality, which especially is well-suited for manipulating artistic images. These methods (Zhang et al. 2020; Zhou et al. 2021; Zhan et al. 2022, 2021; Zhang, Zheng, and Pan 2022) consist of two main phases: cross-domain alignment (align exemplar and conditional input to get aligned representation) and conditional generation (Figure 1(a)).

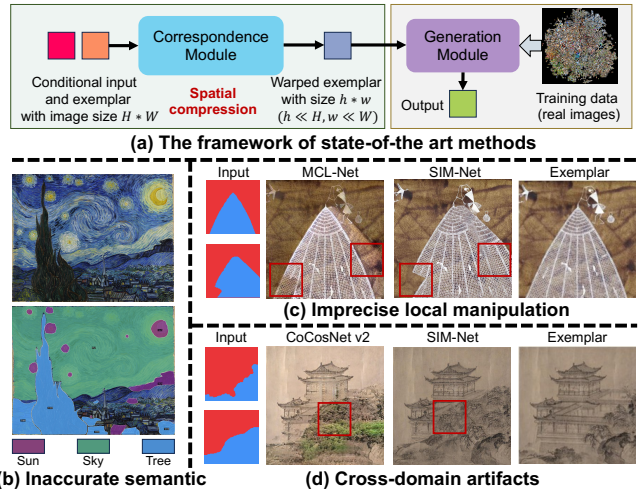

Figure 1: (a) The framework of state-of-the-art exemplar-based image translation methods, such as CoCosNet v2 (Zhou et al. 2021), MCL-Net (Zhan et al. 2022), and MATEBIT (Jiang et al. 2023). (b) These methods require accurate semantic conditional input, while accurate semantic information of artistic images is difficult to extract. (c) The spatial compression in the cross-domain alignment phase leads to imprecise local details. (d) The conditional generation phase might introduce cross-domain artifacts.

However, such methods have following problems when applied to artistic images. **Semantic input.** Existing methods are either based on specific semantic scenarios (*e.g.*, face (Liu et al. 2015; Fan et al. 2022), human pose (Ma et al. 2017; Men et al. 2020; Zhang, Zhan, and Chang 2021), etc.) or require semantic labels (Zhang et al. 2020; Zhou et al. 2021; Zhang, Rao, and Agrawala 2023). However, extracting accurate semantic information from artistic images as input is challenging (Figure 1(b)). **Imprecise control.** Image manipulation requires fine-grained control, while feature compression like multi-level feature pyramids (Zhou et al. 2021; Jiang et al. 2023) leads to imprecise local de-

*These authors contributed equally.
†Corresponding authors.

tails (Figure 1(c)). **Computational efficiency.** Directly generating high-resolution images brings memory overhead at a square level in terms of image size, making it unfeasible. Some methods (Zhan et al. 2022; Jiang et al. 2023) utilize feature pyramids to generate high-resolution images by employing low-resolution images as guidance. However, these methods still exhibit significant computational overhead. **Cross-domain artifacts output.** Existing methods generate images based on the correspondence and the aligned representation, which is typically implemented by unsupervised generative models such as GANs or Diffusion models. However, the correspondence is not always accurate. For inaccurate regions, these methods generate images according to prior knowledge learned from training data which will cause cross-domain artifacts (Figure 1(d)). The style of testing data and training data in artistic image manipulation can hardly be guaranteed to be the same or similar, making it challenging to utilize data priors for generation.

In this paper, we propose an arbitrary **S**tyle **I**mage **M**anipulation **N**etwork (SIM-Net) for exemplar-based artistic image translation, which consists of Mask-based Correspondence Network and Translation Network, similar to the mainstream methods (Figure 1(a)). The Mask-based Correspondence Network takes low level information as input due to the difficulty of extracting semantic information from artistic images, and we use semantic-free masks which have better regional control capability (He et al. 2017). Afterwards, we utilize the Mask-based Correspondence Network to establish the correspondence between two masks guided by the exemplar. These two masks exhibit substantial overlapping regions and few non-overlapping regions. Therefore, we utilize a few number of keypoints to adaptively control different regions for local alignment with low computational overhead implemented by the Local Region Alignment Module. Subsequently, to further obtain the full-resolution correspondence, we dilate these local regions into the global image space and obtain full-resolution warp fields implemented by the Dilating Module. While these several warp fields are full resolution, they focus on respective keypoints regions and provide more precise control. Therefore, we utilize the Generation Network to merge well-controlled regions corresponding to several warp fields to generate precise images. For overlapping regions between masks, we consider using the original exemplar as background estimation to participate in the merging process. Consequently, we propose a *region transportation* strategy to generate images in a self-supervised manner instead of an unsupervised manner implemented by the Image Transport Module, thereby avoiding introducing style features from other domains and cross-domain artifacts. However, this approach introduces splicing artifacts between local regions. To eliminate splicing artifacts, we construct pseudo ground truth to provide both geometrically consistent and spatially consistent supervisory signals implemented by the Texture-Guidance Module, and propose a style self-supervised strategy for training. Our contribution can be summarized as follows:

- We propose a zero-shot arbitrary Style Image Manipulation Network SIM-Net, which does not need to touch any style training data and effectively eliminates cross-domain artifacts.
- We propose a Mask-based Correspondence Network, which ensures a balance between computational efficiency and high resolution, and a *region transportation* strategy for generation in a self-supervised manner.
- Experimental results demonstrate that our method outperforms previous state-of-the-art methods in terms of Style Loss, SSIM, LPIPS, and PSNR.

Related Work

Exemplar-Based Image Translation

Early pioneering works (Zhu et al. 2017; Park et al. 2019; Xu, Zhu, and Wang 2020) attempt to achieve global control over style consistency in generated images by extracting latent codes using style encoder. However, these methods neglected spatial correlations between an input image and an exemplar image, thus failing to produce precise local details. Recently, exemplar-based image translation methods (Zhang et al. 2020; Zhou et al. 2021; Zhan et al. 2021, 2022; Jiang et al. 2023), which leverage an exemplar image to control the style of translated images, establish dense correspondence, and warp the exemplar for generation, have attracted increasing attention. However, as illustrated in Figure 1, these methods are difficult to obtain accurate semantic information in artistic images, and suffer from imprecise local control, substantial computational overhead, and cross-domain artifacts. In contrast, we address these problems.

Low Level Information Guided Methods

The effectiveness of low level priors has been demonstrated in several vision tasks, such as image super-resolution (Li, Zuo, and Loy 2023), image inpainting (Lugmayr et al. 2022), and image restoration (Dogan, Gu, and Timofte 2019). Existing works (Nazeri et al. 2019; Zhang, Rao, and Agrawala 2023) use low level information such as canny edge and sketch for generation. It is worth noting that some of these methods focus on specific semantic scenarios (face (Chen et al. 2020), human pose (Li and Pun 2023), dog (Pan et al. 2023), etc.). Despite their input belonging to low level information, we do not discuss. In contrast to tasks addressed by these methods, we utilize low level information for manipulating artistic images, avoiding cross-domain artifacts associated with high level smenatic information.

Motion Transfer

The motion transfer task aims to drive human body motion based on the given video. These methods (Siarohin et al. 2019, 2021) achieve smooth and natural movement of the human body by capturing the local affine transformations between video frames and a single image to establish global correspondence, which essentially estimates dense optical flow fields between video frames. We refer to the idea of these methods and regard the problem of artistic image manipulation as a mask motion problem.

Proposed Method

Given an input image $x_A \in \mathbb{R}^{H\times W}$ in domain $\mathcal{A}$ and an exemplar artistic image $y_B \in \mathbb{R}^{H\times W\times 3}$ in domain $\mathcal{B}$, our

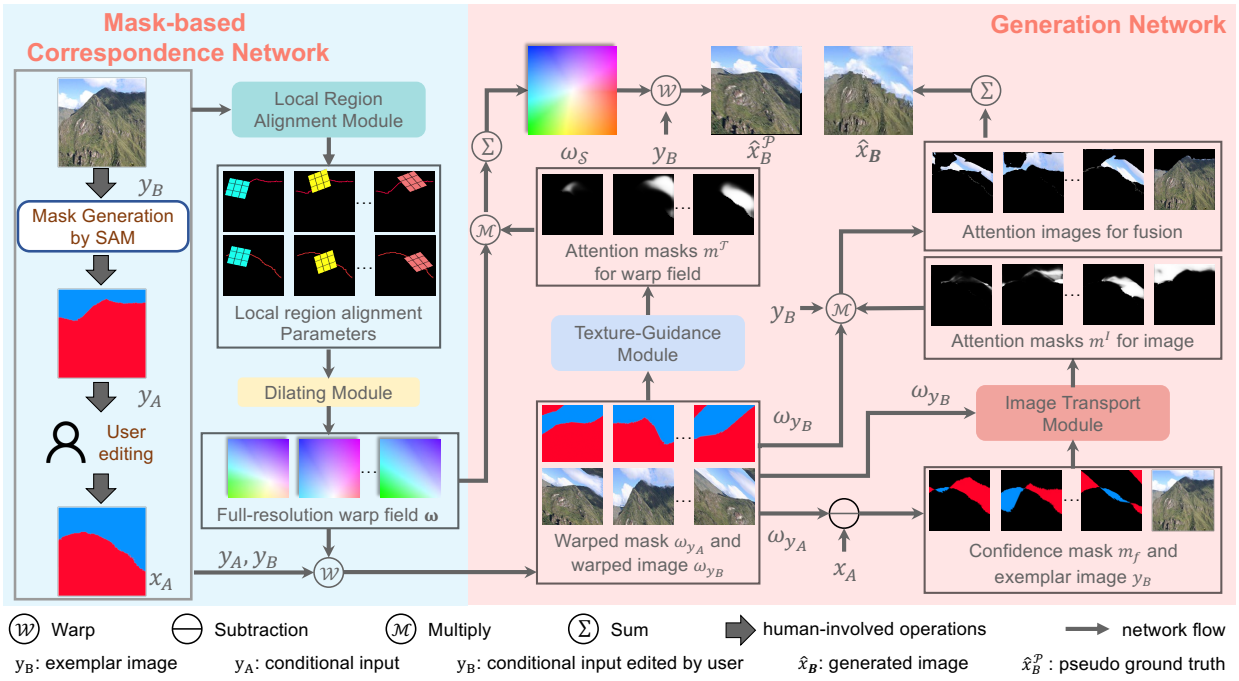

Figure 2: The overall architecture of SIM-Net. The SAM module is used to extract the semantic-free mask of exemplar, denoted as y_A, which is then edited by users to obtain the conditional mask, denoted as x_A. First, the Local Region Alignment Module is used to generate a few number of keypoints that adaptively govern modified regions. Subsequently, the Dilating Module is employed to establish multiple full-resolution corresponding warp fields corresponding to keypoints for global control. Notably, these wrap fields exhibit better control over the region near their corresponding keypoints. Finally, to utilize the characteristics of warp fields, we propose the *region transportation* strategy implemented by the Image Transport Module, utilizing multiple warp fields to construct the generated image, denoted as $\hat{x}_B$. However, $\hat{x}_B$ exhibits splicing artifacts marked by spatial inconsistency. We further design the Texture-Guidance Module to construct the pseudo ground truth, denoted as x_B, serving as a self-supervised signal to eliminate splicing artifacts to ensure spatial consistency.

goal is to generate a image $x_B \in \mathbb{R}^{H\times W\times 3}$ which preserves low level information in x_A and style information in y_B. The exemplar y_B can have arbitrary styles. Figure 2 illustrates an overview of our framework SIM-Net. First, we use semantic-free masks to avoid introducing inaccurate semantic information. Second, we utilize a few number of keypoints to generate full-resolution warp fields, which ensures a balance between computational efficiency and high resolution. Third, we propose a *region transportation* strategy for image generation to avoid cross-domain artifacts.

Mask-Based Correspondence Network

Low Level Input. Previous works (Zhan et al. 2022; Jiang et al. 2023) use an exemplar with a semantic mask to build correspondence. However, the semantic information of artistic images is inaccurate, so we consider using masks with strong region control ability without using semantic information. Segment anything exactly can provide a semantic-free mask $y_A \in \mathbb{R}^{H\times W}$. Users can get x_A by editing the input image y_A, expressed as a binary mask. We consider aligning two masks guided by the exemplar and formulate the problem as: $\mathcal{M}(x_A, y_A|y_B)$.

Local Region Alignment Module. Previous methods use techniques such as feature pyramids (Zhou et al. 2021; Liang, Zeng, and Zhang 2021) to build correspondence which are computationally substantial, and lack precise local control. In artistic image manipulation, two masks primarily maintain unchanged across most regions, with alteration occurring only in specific modified regions (x_A and y_A in Figure 2). Therefore, we propose a strategy that utilizes a few number of keypoints to adaptively govern the modified regions, thereby achieving local alignment. Notably, keypoints have lower computational overhead.

The local alignment between these keypoints is achieved through affine transformation, which is essentially a piecewise linear approximation of the contour so that the original contour becomes the target contour. As the contours are gradually aligned, the corresponding regions are adaptively aligned. In particular, the affine transformation between two masks is essentially a problem of isomorphic change.

We follow the (Siarohin et al. 2019) which is based on a set of learned keypoints with local affine transformations for video sequence tasks to support complex motions, and trained in an unsupervised manner. In our case, we build upon the concept of keypoints to achieve local alignment with lower computational overhead. Multiply y_A and y_B with x_A as input, we can get keypoints $p_1, \cdots, p_K$ and K heatmaps $H^1, \cdots, H^K, s.t. H^k \in [0,1]^{H\times W}$, and further

calculate the affine transformation parameters. Each group of parameters represents image transformation, representing the affine transformation from the virtual reference image R to the input image, specifically formulated as $A^k_{input \leftarrow R}$.

Dilating Module. Although the above module can effectively reduce the computational overhead, we still need global correspondence for accurate manipulation control. Different from the previous method (Siarohin et al. 2019, 2021), which only focuses on local regions, we consider dilating these local regions into the global image space to obtain full-resolution correspondence, denoted as warp fields ω. The implementation is similar to (Siarohin et al. 2019) and under our task can be deduced as follows:

$$\omega^k = \mathcal{D}(x_A, y_A, y_B, A^k) = A^k_{x_A \leftarrow R} \begin{bmatrix} A^k_{y'_A \rightarrow R} \\ 0 \ 0 \ 1 \end{bmatrix}^{-1}, \quad (1)$$

$$y' = y_A \otimes y_B, \quad (2)$$

where $\mathcal{D}$ presents the Dilating Module and y' denotes the combination of y_A and y_B.

As illustrated in Figure 2, ω is then applied to y_A and y_B, generating the warpped image ω_{y_B} and the warpped mask ω_{y_A}. In addition, the full-resolution warp field can be used as a global precise control signal to provide supervisory information for subsequent operations. Through the combination of the above modules, we guarantee the trade-off between computational overhead and full resolution.

Translation Network

Existing methods are trained on realistic images. When dealing with artistic image manipulation, these methods generate images based on the training data prior, introducing cross-domain artifacts. Therefore, we utilize *region transportation* for image generation. Specifically, we propose the Image Transport Module, which can ensure strict control over segmentation and avoid introducing additional content by *region transportation* based on the warp fields. However, *region transportation* will bring splicing artifacts between regions. Furthermore, we propose the Texture-Guidance Module to eliminate artifacts. Details are described below.

Image Transport Module. The previous module yields K sets of full-resolution warp fields, which are derived through keypoint-based local transformations. A single warp field provides more precise control within its own regions, whereas other warp fields provide more control within their respective regions. Therefore, we consider merging them to obtain a more accurate result. We observe that the majority of the two masks remain unchanged and the control capability of these warp fields for unchanged regions is limited. We further incorporate the original exemplar image for background estimation to participate.

Specifically, we propose a strategy to filter out regions with low confidence in each warp field and merge high confidence regions to achieve global control by weighted fusion of attention mechanism. Subsequently, we construct a confidence mask used for filtering out regions, denoted as m_f:

$$\{m^i_f; y_B\} = \{x_A - \omega^i_{y_A}, i = 1, \cdots, K; y_B\}. \quad (3)$$

To achieve more effective control, we compell y_A towards x_A, and strictly ensure the close fit of local contours in a self-supervised manner. Especially, we use K warped masks ω_{y_A}, the original unwarped image for background estimation and K filter masks, and output $K+1$ fused attention maps $m^I_i \in \mathbb{R}^{H \times W}, i = 0, \ldots, K$, which are activated based on softmax to ensure that the sum at each pixel is 1. The original unwarped image is used to keep the content of the unmodified region, which is far from the keypoints:

$$m^I_i(p) = \frac{exp(m^I_i(p))}{\sum_{i=0}^{K} exp(m^I_i(p))}, i = 0, \ldots, K, \quad (4)$$

where $m^I_i(p)$ represents the value of m^I_i at the space coordinate position p. Finally, the image with segmented control is obtained through weighted fusion:

$$\begin{aligned} \hat{x}_A &= \sum_{i=0}^{K} m^I_i \cdot \omega_k(y_A), \\ \hat{x}_B &= \sum_{i=0}^{K} m^I_i \cdot \omega_k(y_B). \end{aligned} \quad (5)$$

As x_A and y_A are paired masks, which provide supervised guidance, we incorporate the BoundaryIoU Loss (Cheng, Schwing, and Kirillov 2021) L_{bound} to minimize the difference between the warped masks ω_{y_A} and x_A, which can get better confidence masks and facilitate the training of the Mask-based Correspondence Network:

$$\mathcal{L}_{\text{bound}} = 1 - \text{BoundaryIoU}(\omega^i_{y_A}, x_A), i = 1, \cdots, K. \quad (6)$$

Texture-Guidance Module. The Image Transport Module brings splicing artifacts between local regions as illustrated by $\hat{x}_B$ in Figure 3. We need ground truth with spatial consistency as supervision, that is, the relative position is determined without splicing artifacts. Subsequently, a trival way is to use the exemplar as ground truth. However, the exemplar is geometrically inconsistent with $\hat{x}_B$, that is, the geometry layout is different, and we lack a means to quantitatively evaluate the similarity between them to facilitate training. Therefore, we further construct a pseudo ground truth that is both spatially consistent and geometrically consistent. Considering warp field ω based on affine transformation parameters, which exhibits strong geometric consistency and spatial consistency, we propose to obtain pseudo ground truth by performing a **single warp** on the original exemplar.

Compared with the fusion of the previous module in the image space, we choose to fuse in the warp field space to obtain a single warp field ω_S by a similar attention module:

$$\omega_{\mathcal{S}}(p) = \sum_{i=1}^{K} m^T_i(p) \cdot \omega_i(p), \quad (7)$$

$$m^T_i(p) = \frac{exp(m^T_i(p))}{\sum_{i=0}^{K} exp(m^T_i(p))}, i = 1, \ldots, K, \quad (8)$$

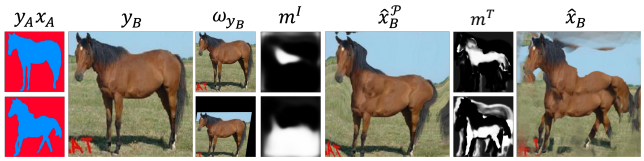

Figure 3: The intermediate results of a training sample in the early epoch. It is evident that $\hat{x}_B^{\mathcal{P}}$ exhibits better geometric consistency and spatial consistency in the early epoch, thanks to the geometric consensus achieved through the warp fields. However, the layout of $\hat{x}_B^{\mathcal{P}}$ is less consistent with x_A. Additionally, it can be observed that $\hat{x}_B$ demonstrates improved semantic consistency as a result of the fusion of several candidates. However, this process can introduce fusion and splicing artifacts that may disrupt the geometric consistency and spatial consistency. The results provide visual evidence of the trade-off between geometric consistency, spatial consistency, and semantic consistency in the intermediate results during the early training epochs.

where $m_i^T(p)$ represents the value of m_i^T at the space coordinate position p.

The pseudo ground truth $\hat{x}_A^{\mathcal{P}}$ and $\hat{x}_B^{\mathcal{P}}$ are obtained by swaping y_A and y_B through a single warp field $\omega_{\mathcal{S}}$. We further propose the style self-supervision strategy, which constrains the output $\hat{x}_B$ to maintain style consistency with the pseudo ground truth $\hat{x}_B^{\mathcal{P}}$, as measured by the Style Loss (Gatys, Ecker, and Bethge 2016). We use Contextual Loss (Mechrez, Talmi, and Zelnik-Manor 2018) to minimize the style difference. The contextual loss is computed by:

$$\mathcal{L}_{\text{context}} = \sum_l w_l[-log(CX(\phi^l(\hat{x}_B), \hat{x}_B^{\mathcal{P}}))], \qquad (9)$$

where w is used to adjust the weights between different network layers to balance the loss of each layer, $CX(x, y)$ represents the similarity between feature maps.

Compared to MSE Loss and SSIM Loss, which have stricter structural constraints, our method, as mentioned before, ensures structural consistency, and can more effectively enforce constraints on color and texture. In addition, compared with directly using the exemplar y_B as style information guidance, such a strategy can further reduce the impact of image semantic content on style feature extraction, and pay more attention to the differences brought by style.

To facilitate training and ensure the generation of images with geometric consistency and spatial consistency, we use the cycle loss function $\mathcal{L}_{\text{cyc}}$. Specifically, through swapping x_A and y_A, we utilize $\hat{x}_B^{\mathcal{P}}$ as the exemplar and generates the pseudo original image $\hat{y}_B$. The cycle loss function minimizes the ℓ_1 distance between y_B and $\hat{y}_B$. It is worth noting that both the Image Transport Module and Texture-Guidance Module are trained based on the same Mask-based Correspondence Network, thereby constraining each other.

Additional Loss Functions

We use Equivariance constraint Loss $\mathcal{L}_{\text{eq}}$ (Siarohin et al. 2019) to ensure the stable training of the Local Region Alignment Module, use Perceptual Loss $\mathcal{L}_{\text{perc}}$ to constrain the image structure, and use Conditional alignment Loss $\mathcal{L}_{\text{mask}}^I$ and $\mathcal{L}_{\text{mask}}^T$ to ensure consistency of mask structure.

Total Loss. Our framework is end-to-end optimized to jointly achieve zero shot arbitrary style image manipulation and alternate between optimizing Mask-based Correspondence Network, Image Transport Module, and Texture-Guidance Module using the aforementioned loss functions. The overall objective function of proposed framework is:

$$\begin{aligned}\mathcal{L}_{\text{total}} =& \lambda_1\mathcal{L}_{\text{eq}} + \lambda_2\mathcal{L}_{\text{perc}} + \lambda_3\mathcal{L}_{\text{context}} + \lambda_4\mathcal{L}_{\text{bound}} \\ &+ \lambda_5(\mathcal{L}_{\text{mask}}^I + \mathcal{L}_{\text{mask}}^S) + \lambda_6\mathcal{L}_{\text{rec}} + \lambda_7\mathcal{L}_{\text{cyc}},\end{aligned} \qquad (10)$$

where λ is weighting parameter.

Inference. The output of $\hat{x}_B$ is taken as the manipulation result. Note that manipulating several regions could be achieved by sequentially feeding the output to the model.

Experiments

Implementation details. The learning rate for the framework is $2e-4$, and we use the Adam solver with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ for the optimization. K is set to be 10. Unless otherwise specified, the resolution of generated images for translation tasks is 256×256 for fair comparison. The experiments are conducted using 4 TITAN Xp GPUs.

Training dataset. Our training data is composed of 1,000 faces images from CelebAMask-HQ (Lee et al. 2020), 227 horses images from Weizmann Horse Database (Borenstein, Sharon, and Ullman 2004), and 696/573 mountains/buildings images from Intel Image Classification. The size is 2,496, which is significantly smaller as compared with other GAN-based methods (*e.g.*, 20,210 images from ADE20k (Zhou et al. 2017) used by (Zhang et al. 2020; Zhou et al. 2021; Zhan et al. 2021)).

Testing dataset. We collect 237 artistic images with 10 styles for evaluation, including painting styles of abstract (22 images, abbr. as AB), cubism (29 CU), India (22 IN), Chinese (21 CH), expression (21 EX), Japanese (32 JA), modernism (21 MO), impressionism (24 IM), photorealistic (24 PH), and surrealism (21 SU).

Evaluation Metrics. Since our method is not a GAN-based approach, and our training data and generated images are with quite different domains, metrics such as FID (Heusel et al. 2017) that aims to compute the distance between Gaussian fitted feature distributions of realistic and generated images, and SWD (Karras et al. 2018) that attempts to measure the Wasserstein distance between the distribution of realistic images and generated images are not suitable for evaluation. We present quantitative evaluation from several directions. (1) Style Loss (Gatys, Ecker, and Bethge 2016), which has been used to measure the style similarity (Wu et al. 2021; Wang et al. 2020), is the mean-squared distance of features extracted from a pre-trained VGG model (Simonyan and Zisserman 2014). (2) SSIM (Wang, Simoncelli, and Bovik 2003), PSNR (Huynh-Thu and Ghanbari 2008) and LPIPS (Zhang et al. 2018) are used to measure the image quality.

	AB, CU, IN (73)		CH, EX, JA, MO (95)		IM, PH, SU (69)		ALL (237)				
	Style Loss (ROI)	SSIM (ROU)	Style Loss (ROI)	SSIM (ROU)	Style Loss (ROI)	SSIM (ROU)	Style Loss (ROI)	SSIM (ROU)	Style Loss (Whole)	PSNR (ROU)	LPIPS (ROU)
INADE	21.24	0.28	10.36	0.34	5.92	0.34	12.29	0.32	32.02	10.605	0.574
PZ20	6.27	0.61	3.77	0.67	2.43	0.67	4.12	0.65	14.04	20.449	0.234
PMD	3.35	0.47	2.34	0.53	1.49	0.56	2.39	0.52	3.46	17.638	0.176
LY21	3.41	0.58	2.36	0.63	1.61	0.63	2.45	0.61	5.29	19.366	0.151
CoCosNet	4.95	0.57	3.53	0.61	1.96	0.60	3.49	0.59	8.10	18.361	0.249
UNITE	3.07	0.76	2.10	0.81	1.40	0.81	2.18	0.79	3.19	22.272	0.096
CoCosNet v2	5.02	0.61	3.48	0.65	2.21	0.64	3.58	0.63	8.26	18.821	0.243
MCL-Net	2.83	0.84	1.85	0.87	1.33	0.86	2.01	0.86	3.05	25.688	0.087
DynaST	4.87	0.59	3.86	0.65	2.03	0.71	3.64	0.65	7.67	19.772	0.194
MATEBIT	4.12	0.64	3.51	0.69	1.84	0.73	3.21	0.69	7.27	20.541	0.173
SIM-Net	**2.20**	0.95	**1.59**	0.95	1.07	0.96	1.62	0.95	**1.31**	29.570	0.031
SIM-Net *w* DM	2.23	**0.97**	1.61	**0.96**	**1.02**	**0.97**	**1.63**	**0.97**	1.34	**30.124**	**0.028**

Table 1: Quantitative comparison in terms of Style Loss, SSIM, PSNR, and LPIPS. DM denotes as diffusion model.

Overall Performance

Comparison with state-of-the-art. We compare qualitative and quantitative performance with state-of-the-art image manipulation methods, including two latent vector editing approaches INADE (Tan et al. 2021) and swapping autoencoder (Park et al. 2020), two semantic correspondence approaches PMD (Li et al. 2021) and learning to warp (Liu, Yang, and Hall 2021) (for general neural style transfer), and exemplar-based image translation methods CoCosNet (Zhang et al. 2020), UNITE (Zhan et al. 2021), CoCosNet v2 (Zhou et al. 2021), MCL-Net (Zhan et al. 2022), DynaST (Liu et al. 2022) and MATEBIT (Jiang et al. 2023). Note that these methods as well as SIM-Net are trained on realistic images. Furthermore, we propose an alternative version, SIM-Net *w* DM, which combines our method with a pretrained diffusion model. More details can see (Guo et al. 2024).

Quantitative evaluation. The quantitative comparison results are shown in Table 1. As can be observed from Table 1, comparison methods (*e.g.*, INADE, UNITE) produce less competitive results with styles of AB, CU, and IN, while achieving better results for those of IM, PH, and SU. Because these works are more likely to glean visual features of realistic training data and the former/latter styles are less/-more realistic. Compared with existing methods, our model achieves competitive performance across all style images.

Qualitative evaluation. Figure 5 illustrates images generated by different state-of-the-art methods. Previous methods present local inconsistency and cross-domain artifacts due to the training data prior and space compression. In contrast, our results preserve consistent structure with edited mask, consistent style with exemplar image, and present consistent appearance with similar regions with the exemplar image.

Inference Speed and Computational Efficiency. To prove the advantages of our method's low computation and full precision, we choose to conduct comparative experiments with state-of-the-art methods on 20 high-resolution images. For a fair comparison, we choose to use the results without a diffusion module. As shown in Table 2, our method achieves a balance of full resolution and low computation, and maintains fast inference speed. As illustrated in Figure 4, our method achieves precise manipulation and guarantees high resolution. Besides, when there is a significant difference between the edited mask and the input, our method can address it through multi-step operations with fast response.

Figure 4: Qualitative comparison of our method and state-of-the-art methods in terms of high-resolution.

	color	texture	fake detection	time	memory
CoCosNet	0.752	0.536	0.42	0.716	162.9
CoCosNet v2	0.795	0.616	0.38	3.248	**59.3**
MCL-Net	0.884	0.758	0.25	0.574	215
MATEBIT	0.847	0.771	0.31	0.301	114.7
SIM-Net	**0.925**	**0.814**	**0.17**	**0.126**	89.6

Table 2: Quantitative evaluation in terms of style relevance (color and texture), fake detection, time, and memory.

Effectiveness of Eliminating Cross-Domain Artifacts. The previous style loss demonstrated the overall difference between images, and we further demonstrate that our method does not introduce cross-domain artifacts. Specifically, the style relevance is proved through the three dimensions of color, texture and fake detection. We compare the cosine similarities between low level features to measure color and texture relevance and use mvssnet++ (Chen et al. 2021; Dong et al. 2022) for fake detection with a threshold of 0.04. Table 2 shows that our method has clear advantages over existing methods. Figure 6 illustrates that our method has relatively fewer artifacts of manipulation.

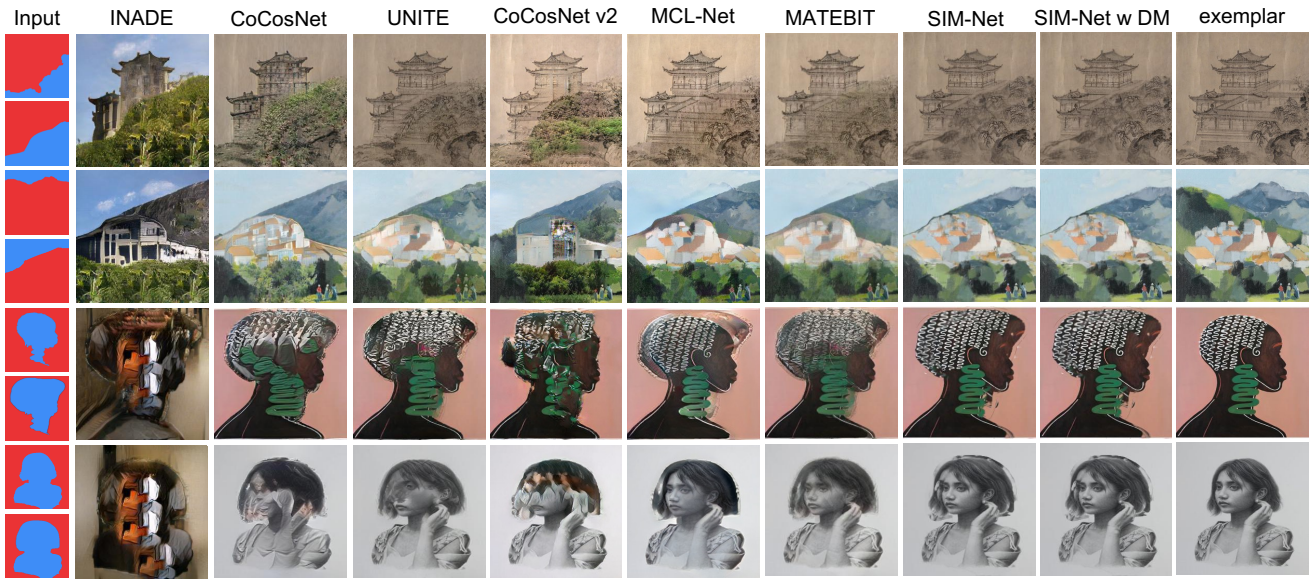

Figure 5: Visual qualitative comparison with state-of-the-art methods. It can be seen that our method has no cross-domain artifacts and fine details without blurring.

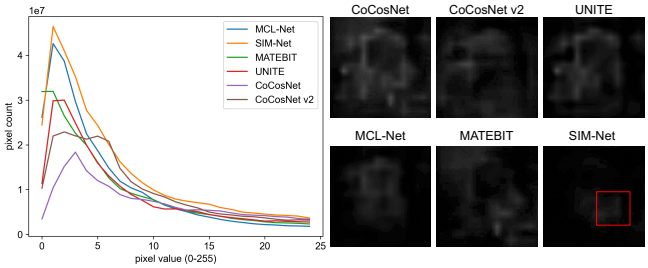

Figure 6: Qualitative and quantitative comparison for manipulation artifacts (row 1 in Figure 5). The horizontal axis represents pixel values (0-255), and the vertical axis represents the pixel count. The red box has slight tampering marks.

	Style Loss	SSIM	PSNR	LPIPS
w/o $\mathcal{L}_{\text{bound}}$	1.78	0.79	28.114	0.057
w/o $\mathcal{L}_{\text{context}}$	2.24	0.92	25.126	0.174
w/o $\mathcal{L}_{\text{cyc}}$	1.92	0.72	26.358	0.122
w semantic mask	**1.09**	0.96	29.963	0.033
w/o Style Self-Supervision	10.16	0.76	22.440	0.184
w generation model	8.19	0.62	20.374	0.151
Full model: SIM-Net	1.34	**0.97**	**30.124**	**0.028**

Table 3: Ablation study of SIM-Net.

Figure 7: Introduce GAN loss which makes the representation effect insensitive to locality. Note that conditional semantic masks are enhanced to highlight their differences for better visualization.

Ablation Study

Table 3 shows our method (full model) demonstrates best performance. Specifically, (1) remove $\mathcal{L}_{\text{bound}}$ to validate structural consistency (see SSIM); (2) remove $\mathcal{L}_{\text{context}}$ to validate style minimization (see Style Loss); (3) remove $\mathcal{L}_{\text{cyc}}$ to validate the importance for the mask-based correspondence network; (4) replace semantic-free mask with semantic mask to explore the impact of semantic information; (5) remove style self-supervision strategy to validate the necessity on generative quality, which tends to produce results with blurry artifacts; (6) replace our generation network with GANs to validate its effectiveness.

Unnecessity of GAN Loss. To prove the Unnecessity of GAN Loss, we have tried using adversarial loss by adding an additional discriminator (the same network architecture in (Zhan et al. 2022)). As can be observed from Figure 7, although the cross-domain artifacts are suppressed by our method, the results are insensitive to local manipulation.

Conclusion

This paper presents SIM-Net, a novel zero-shot image manipulation framework. The proposed scheme is capable of handling diverse styles of artistic images without relying on large-scale image training. Through quantitative and quality experiments with state-of-the-art methods, our approach demonstrates effective image processing capabilities.

Limitations. Our method is less creative, and is suitable for precisely controlled scenarios. Our method affects artistic aesthetics, such as brush strokes, which is an advanced issue in the application of AI to art, requiring prior knowledge from domain experts (He et al. 2018).

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.U20A20220), the grants from Key R&D Program of Zhejiang (no. 2022C01048), Key Program of National Natural Science Foundation of China (62334014).

References

Borenstein, E.; Sharon, E.; and Ullman, S. 2004. Combining top-down and bottom-up segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*

Chen, S.-Y.; Su, W.; Gao, L.; Xia, S.; and Fu, H. 2020. DeepFaceDrawing: Deep generation of face images from sketches. *ACM Trans. Graph.*

Chen, X.; Dong, C.; Ji, J.; Cao, J.; and Li, X. 2021. Image manipulation detection by multi-view multi-scale supervision. In *Int. Conf. Comput. Vis.*

Cheng, B.; Schwing, A.; and Kirillov, A. 2021. Per-pixel classification is not all you need for semantic segmentation. *Adv. Neural Inform. Process. Syst.*

Dogan, B.; Gu, S.; and Timofte, R. 2019. Exemplar guided face image super-resolution without facial landmarks. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*

Dong, C.; Chen, X.; Hu, R.; Cao, J.; and Li, X. 2022. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Trans. Pattern Anal. Mach. Intell.*

Fan, D.-P.; Huang, Z.; Zheng, P.; Liu, H.; Qin, X.; and Van Gool, L. 2022. Facial-sketch synthesis: a new challenge. *Machine Intelligence Research*.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Guo, W.; Zhang, Y.; Ma, D.; and Zheng, Q. 2024. Learning to Manipulate Artistic Images. *arXiv preprint arXiv:2401.13976*.

He, B.; Gao, F.; Ma, D.; Shi, B.; and Duan, L.-Y. 2018. Chipgan: A generative adversarial network for chinese ink wash painting style transfer. In *ACM Int. Conf. Multimedia*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Int. Conf. Comput. Vis.*

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inform. Process. Syst.*

Huynh-Thu, Q.; and Ghanbari, M. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics letters*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Jiang, C.; Gao, F.; Ma, B.; Lin, Y.; Wang, N.; and Xu, G. 2023. Masked and Adaptive Transformer for Exemplar Based Image Translation. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *Int. Conf. Learn. Represent.*

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Li, H.; and Pun, C.-M. 2023. CEE-Net: complementary end-to-end network for 3D human pose generation and estimation. In *AAAI*.

Li, X.; Fan, D.-P.; Yang, F.; Luo, A.; Cheng, H.; and Liu, Z. 2021. Probabilistic Model Distillation for Semantic Correspondence. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Li, X.; Zuo, W.; and Loy, C. C. 2023. Learning Generative Structure Prior for Blind Text Image Super-Resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Liang, J.; Zeng, H.; and Zhang, L. 2021. High-resolution photorealistic image translation in real-time: A laplacian pyramid translation network. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. *Adv. Neural Inform. Process. Syst.*

Liu, S.; Ye, J.; Ren, S.; and Wang, X. 2022. Dynast: Dynamic sparse transformer for exemplar-guided image generation. In *Eur. Conf. Comput. Vis.*

Liu, X.-C.; Yang, Y.-L.; and Hall, P. 2021. Learning To Warp for Style Transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Int. Conf. Comput. Vis.*

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Ma, L.; Jia, X.; Sun, Q.; Schiele, B.; Tuytelaars, T.; and Van Gool, L. 2017. Pose guided person image generation. *Adv. Neural Inform. Process. Syst.*

Mechrez, R.; Talmi, I.; and Zelnik-Manor, L. 2018. The contextual loss for image transformation with non-aligned data. In *Eur. Conf. Comput. Vis.*

Men, Y.; Mao, Y.; Jiang, Y.; Ma, W.-Y.; and Lian, Z. 2020. Controllable person image synthesis with attribute-decomposed gan. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.; and Ebrahimi, M. 2019. Edgeconnect: Structure guided image inpainting using edge prediction. In *Int. Conf. Comput. Vis. Worksh.*

Pan, X.; Tewari, A.; Leimkühler, T.; Liu, L.; Meka, A.; and Theobalt, C. 2023. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Park, T.; Zhu, J.-Y.; Wang, O.; Lu, J.; Shechtman, E.; Efros, A. A.; and Zhang, R. 2020. Swapping Autoencoder for Deep Image Manipulation. In *Adv. Neural Inform. Process. Syst.*

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. First order motion model for image animation. *Adv. Neural Inform. Process. Syst.*

Siarohin, A.; Woodford, O. J.; Ren, J.; Chai, M.; and Tulyakov, S. 2021. Motion representations for articulated animation. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tan, Z.; Chai, M.; Chen, D.; Liao, J.; Chu, Q.; Liu, B.; Hua, G.; and Yu, N. 2021. Diverse Semantic Image Synthesis via Probability Distribution Modeling. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Wang, H.; Li, Y.; Wang, Y.; Hu, H.; and Yang, M.-H. 2020. Collaborative distillation for ultra-resolution universal style transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*.

Wu, X.; Hu, Z.; Sheng, L.; and Xu, D. 2021. StyleFormer: Real-Time Arbitrary Style Transfer via Parametric Style Composition. In *Int. Conf. Comput. Vis.*

Xu, S.; Zhu, Q.; and Wang, J. 2020. Generative image completion with image-to-image translation. *Neural Computing and Applications*.

Zhan, F.; Yu, Y.; Cui, K.; Zhang, G.; Lu, S.; Pan, J.; Zhang, C.; Ma, F.; Xie, X.; and Miao, C. 2021. Unbalanced Feature Transport for Exemplar-based Image Translation. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Zhan, F.; Yu, Y.; Wu, R.; Zhang, J.; Lu, S.; and Zhang, C. 2022. Marginal contrastive correspondence for guided image generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Zhang, C.; Zhan, F.; and Chang, Y. 2021. Deep monocular 3d human pose estimation via cascaded dimension-lifting. *arXiv preprint arXiv:2104.03520*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Int. Conf. Comput. Vis.*

Zhang, P.; Zhang, B.; Chen, D.; Yuan, L.; and Wen, F. 2020. Cross-domain correspondence learning for exemplar-based image translation. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Zhang, Y.; Zheng, Q.; and Pan, G. 2022. EasyPainter: Customizing Your Own Paintings. In *CAAI International Conference on Artificial Intelligence*.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Zhou, X.; Zhang, B.; Zhang, T.; Zhang, P.; Bao, J.; Chen, D.; Zhang, Z.; and Wen, F. 2021. CoCosNet v2: Full-Resolution Correspondence Learning for Image Translation. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Int. Conf. Comput. Vis.*