SwinIT: Hierarchical Image-to-Image Translation Framework Without Cycle Consistency

Jin Liu, Huiyuan Fu, *Member, IEEE*, Xin Wang, *Senior Member, IEEE*, and Huadong Ma, *Fellow, IEEE*

***Abstract*— Image-to-image (I2I) translation often requires establishing cycle consistency between the source and the translated images across different domains. However, cycle consistency requires redundant reconstruction, and is too restrictive to satisfy the bijection assumption between the two domains. In this paper, we propose SwinIT, a hierarchical Swin-transformer I2I Translation framework without using cycle consistency. Specifically, we carefully design symmetrical encoders for content and style flows, then explore newly proposed adaptive denormalization and normalization strategies. This framework can effectively capture and fuse content and style representations in a coarse-to-fine manner, ensuring our method achieves high performance without cycle consistency. Guided by element-wise feature adaptive denormalization, our model focuses on preserving semantic structure information. Due to the semantic mismatch between unpaired source and exemplar images, we introduce cross-attention adaptive instance normalization to help achieve better alignment. However, because the original optimization objective lacks direct supervision to preserve high-frequency information, rich edge details are lost during the translation. We propose a wavelet transformation matching loss to recover the details by converting the image into multi-frequency parts. We validate our proposed method in various I2I translation tasks, including arbitrary style transfer, multi-modal image synthesis, and semantic image synthesis, demonstrating its effectiveness in both qualitative and quantitative evaluations.**

***Index Terms*— Transformer, wavelet transformation, image-to-image translation.**

I. Introduction

IMAGE-TO-IMAGE (I2I) translation has drawn a lot of attention in recent years, which has been broadly applied in various computer vision tasks such as synthesis [1], [2], [3], [4], super-resolution [5], [6], [7], style transfer [8], [9], colorization [10], [11], etc. I2I focuses on solving a disentanglement problem that retains the structure or content separated from the image while changing the appearance or style.

In a general situation, when training only unpaired data, the source content is preserved by using cycle consistency scheme [12], [13], [14] while the target style is transferred by using adversarial and perceptual losses [15]. Cycle consistency aims to learn an inverse mapping that translates the output from the target domain back to the input, while minimizing the per-pixel distance between the generated sample and ground truth. Furthermore, MUNIT [16] and DRIT++ [17] learn a shared intermediate latent space to enable multimodal and multiple domain synthesis. They use cycle consistency constraints in an image or latent space.

To ensure cycle consistency, two domains are required to satisfy the bijective assumption, which is too restrictive and often causes the steganography phenomenon [18] that the generated images are blurry and lose texture details. In addition, the complicated and tedious reconstruction path increases the computational complexity of the gradient optimization.

Alternative or complementary solutions have been used to confront these challenges [18], [19], [20]. For instance, Park et al. [19] propose a straightforward method to preserve the correspondence in the content space. They use multilayer and patchwise contrastive loss between the corresponding input and output patches to replace cycle consistency loss. To satisfy the cycle consistency, Gao et al. [18] apply the wavelet-based skip connection component to explicitly represent omitted high-frequency information. Nevertheless, specialized structures for feature space sampling or representation are required to maintain spatial coherence and resolution. Instead, we aim to design a succinct framework without specially designed components to facilitate training for I2I translation tasks.

We propose to build a GAN-based framework that uses Transformers for I2I translation without resorting to cycle consistency. Recently, vanilla Transformers have achieved great success in a broad range of visual tasks [9], [21], [22], [23], [24], [25] (i.e., image classification, detection, segmentation, and image generation). Among them, Swin Transformer [22] takes advantage of the use of both CNN and Transformer, and has shown great promise. More specifically, it not only employs a window-based local attention mechanism but also models the long-range dependency with the shifted window scheme. Hence, we propose SwinIT, a generative framework that uses symmetrical encoders for content and style flows.

Manuscript received 11 June 2023; revised 5 October 2023 and 25 December 2023; accepted 29 December 2023. Date of publication 15 January 2024; date of current version 3 July 2024. This work was supported in part by NSFC under Grant 62272059, in part by the National Key Research and Development Program of China under Grant 2023YFF0904800, in part by the Beijing Nova Program under Grant 20230484406, in part by the Innovation Research Group Project of NSFC under Grant 61921003, and in part by the 111 Project under Grant B18008. This article was recommended by Associate Editor L. Nie. *(Corresponding author: Huiyuan Fu.)*

Jin Liu, Huiyuan Fu, and Huadong Ma are with the Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: ljin@bupt.edu.cn; fhy@bupt.edu.cn; mhd@bupt.edu.cn).

Xin Wang is with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794 USA (e-mail: x.wang@stonybrook.edu).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TCSVT.2024.3353932.

Digital Object Identifier 10.1109/TCSVT.2024.3353932

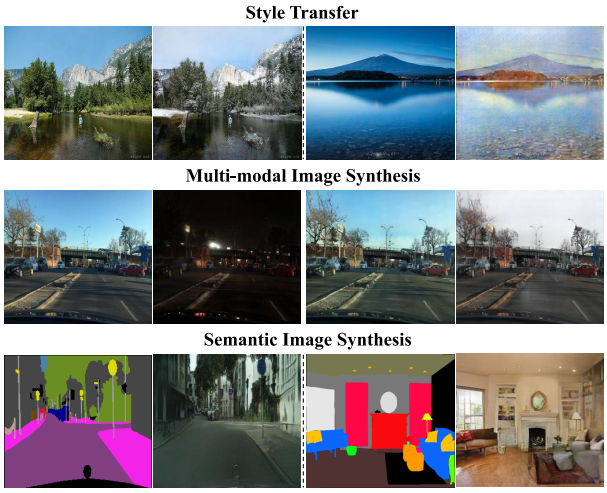

Fig. 1. Image samples generated by our SwinIT for various image-to-image translation tasks without cycle consistency.

It is versatile for various I2I translation tasks, as shown in Fig. 1. We leverage Swin Transformers as the basic building block and take advantage of the hierarchical structure to characterize different feature scales. Without the supervision of cycle consistency, we cannot guarantee that the generated target image has a similar distribution or characteristics to both the source domain (content) and the reference domain (style). Therefore, we consider both the content and style encoding flows that learn disentangled representations at multi-scale feature levels and adaptively fuse them in a coarse-to-fine manner. Specifically, since the normalization layers in the network tend to clear the semantic information contained in the input image, we propose element-wise Transformer Adaptive DEnormalization (TADE), a conditional normalization strategy that modulates the features through adaptive and learned transformation and can help preserve the content spatial structure. In real scenarios, unpaired one-to-one image translation often suffers from the semantic mismatch and the mutual incompatibility between source and reference examples. Previous style normalization strategies like AdaIN [26] or SPADE [1] can merely generate images with overall similarities to the reference. To address this issue, we introduce Cross-Attention style Adaptive Instance Normalization (CA-AdaIN), which learns the correlation between the source content and exemplar style to facilitate better alignment. It achieves precise style representations through content-adaptive matching.

Nonetheless, we observe that rich details (e.g., high-frequency edges and textures) are lost and unpleasant artifacts appear during translation. The traditional optimization objective [1], [27], [28] mainly ensures high-level style and content feature consistency, and performs image-level reconstruction. These strategies are either coarse-grained or sensitive to small perturbations. In the field of computer vision, wavelet transformation is known to be good at decomposing images into different frequency bands to capture rich details. To capture the rich image details after the translation, we convert the generated image to multi-frequency parts with the supervision of a real image. The wavelet transform constraint is introduced to help the network preserve the rich texture details and suppress the artifacts. The main contributions of this work are summarized as follows:

- We introduce SwinIT, a hierarchical framework without cycle consistency constraints that can be effectively used in various I2I translation tasks.
- We design multi-scale feature denormalization (TADE) and normalization (CA-AdaIN) schemes to capture and fuse content and style information from coarse to fine.
- We propose a novel wavelet transform matching loss to help our model recover rich details by transforming images into high-frequency domains.
- We conduct extensive qualitative and quantitative experiments to demonstrate the superiority of our proposed approach on several standard benchmarks.

Equipped with the above techniques, SwinIT yields visually pleasing outputs and substantially outperforms several classic convolutional baselines.

II. Related Work

A. Generative Adversarial Network

Generative adversarial networks (GAN) [29] are widely employed in diverse high-quality generating tasks such as images and videos [11], [15], [30], [31], [32]. The framework of GAN consists of a generator and a discriminator. The generator tries to generate fake but realistic images to fool the discriminator, and the discriminator attempts to distinguish whether a generated image is real or fake. Various GAN-based models are proposed to optimize training stability. CGAN [33] can not only output realistic images but also meet the conditional description. Wasserstein GAN (WGAN) [34] uses Wasserstein distance to solve the gradient vanishing and mode collapse problems. Deep convolutional GAN (DCGAN) [35] adopts a fully convolutional architecture, using strided convolution and batch normalization for more stable training and better performance.

Different from the above methods, in this work, we exploit hierarchical encoders and decoders in our generator to capture and fuse different representations in a coarse-to-fine manner. In addition, we apply patch-based training to discriminate images at different scales, further improving the robustness of our method for I2I translation tasks.

B. Image-to-Image Translation

Image-to-image (I2I) translation has made great progress in recent years as it is widely used in super-resolution [30], [36], [37], [38], [39], [40], colorization [11], [41], [42], [43], inpainting [44], [45], [46] and style transfer [13], [15], [20], [32], [47], [48], [49], etc.

We classify I2I translations into supervised and unsupervised categories. Supervised methods [15], [50], [51] require pixel-aligned paired training samples. They mainly aim to turn sketches or semantic layouts into photorealistic images. Park et al. [1] introduce GauGAN with spatially adaptive normalization for image synthesis, keeping the effective semantic meaning. In contrast, unsupervised methods generally train unpaired data with the constraint of cycle consistency [12],

[13], [52], deep perceptual features [53] or contrastive learning [19]. They aim to transfer content images in a given reference style while preserving the content representations. AdaIN [26] is an effective normalization strategy widely used to accomplish arbitrary style transfer tasks.

Limited to two-domain mapping translation, recently researchers have made great efforts to explore deeper into multi-modal translation [16], [54], [55] and multi-domain translation [17], [28], [56] for generation diversity. For example, MUNIT [16] trains diverse image translation maps by disentangling domain-invariant content space and domain-specific style codes. StarGANs [2], [57] learn a mapping between different visual domains while satisfying the diversity of generated images and scalability across multiple domains. Wang et al. [58] propose a shared knowledge module to learn the common information among multi-domain pairs.

However, existing I2I translation methods require complex settings such as auxiliary networks, loss functions (e.g., cycle consistency), and supervised domain labels. The different needs of unsupervised and supervised settings force previous approaches to utilize custom modules, which leads to a degradation in quality. In this work, we propose a succinct framework to adapt to different tasks and improve scalability and robustness. Our model achieves outstanding translation performance based on multi-scale feature denormalization and normalization operations from coarse to fine that adapt to various tasks.

C. Transformers in Computer Vision

Transformer is a type of neural network based on the multi-head self-attention mechanism. Due to the strong representation capabilities in the NLP field, researchers attempt to demonstrate the great potential of Transformers in a broad range of discriminative and generative vision tasks [21], [22], [23], [24], [59]. For instance, Dosovitskiy et al. [21] propose a seminal model called ViT, which is a pure Transformer architecture for image classification by treating an image as a sequence of 16 × 16 visual words. Han et al. [60] introduce the first Transformer model IPT for multiple low-level vision tasks by using large pre-training datasets. Recently, a few works [61], [62] propose Transformer-based GANs to achieve image synthesis. As a pure Transformer-based GAN architecture, Jiang et al. [61] build a memory-friendly generator and a multi-scale discriminator for image generation, and study new training recipes to achieve competitive performance. Furthermore, inspired by [25] and [63] presents InstaFormer to integrate global- and instance-level information, but only applies ViT structure in encoder blocks. Reference [9] completes a single-style transfer task with Transformers.

To sum up, existing methods lack the scalability to adapt to different tasks in unsupervised and supervised settings. In this work, we introduce a versatile Transformer-based framework, dubbed SwinIT for I2I translation. The dual-stream architecture makes it suitable to serve as general-purpose backbones. SwinIT can synthesize fine structures using a cascade of Swin Transformer-based blocks, leading to comparable quality as the leading ConvNets in realistic synthesis.

III. Framework of SwinIT

We attribute the good performance and robustness of our framework on various translation tasks to the following three aspects: 1) The disentangled content and style representation by Transformer-based encoders. 2) The content and style features are injected into multi-scale feature levels for strong training stability. 3) Concise and effective optimization objectives to fit diverse translation tasks.

A. Transformer-Based GAN Architecture

Benefiting from the outstanding performance of networks based on Swin Transformer [22] in discriminative and generative tasks [62], [64], [65], we build a **Swin**-transformer based **I**mage-to-image **T**ranslation (SwinIT) framework, as shown in Fig. 2. It contains three components, including symmetrical content and style encoders, a generator, and the omitted discriminator. Each component consists of several Transformer blocks.

1) Content and Style Encoder: We take two symmetrical networks to disentangle content representations from original images and style representations from reference images on different levels. We construct the submodules of encoders based on the Swin Transformer blocks [22] which compute multi-head self-attention (MSA) locally in non-overlapping windows. To achieve an enlarged receptive field, we use a parallelized double attention mechanism [62] which allows a single block to simultaneously attend to the context of the local and shifted windows. Specifically, let $\boldsymbol{X^i} \in \mathcal{R}^{H\times W\times C}$ denotes the input feature map of layer i. We split h attention heads into two groups: the first half $\boldsymbol{X_w^i} \in \mathcal{R}^{H\times W\times \frac{C}{2}}$ of heads perform the regular window attention while the second half $\boldsymbol{X_{sw}^i} \in \mathcal{R}^{H\times W\times \frac{C}{2}}$ compute the shifted window attention, both of whose results are further concatenated to form the output. The double attention is formulated as:

$$\begin{aligned}\hat{\boldsymbol{X}}_{\boldsymbol{w}}^{\boldsymbol{i}} &= \text{W-MSA}(\text{LN}(\boldsymbol{X}_{\boldsymbol{w}}^{\boldsymbol{i}})),\\ \hat{\boldsymbol{X}}_{\boldsymbol{sw}}^{\boldsymbol{i}} &= \text{SW-MSA}(\text{LN}(\boldsymbol{X}_{\boldsymbol{sw}}^{\boldsymbol{i}})),\\ \boldsymbol{X}_{\boldsymbol{double\text{-}attn}} &= \text{MLP}(\text{Concat}(\hat{\boldsymbol{X}}_{\boldsymbol{w}}^{\boldsymbol{i}}, \hat{\boldsymbol{X}}_{\boldsymbol{sw}}^{\boldsymbol{i}})) + \boldsymbol{X^i}, \end{aligned} \tag{1}$$

where W-MSA and SW-MSA denote the window-based multi-head self-attention under regular and shifted window partitioning, respectively, LN stands for layer normalization, and MLP denotes one fully connected layer.

Relative positional encoding (RPE) [22] is adopted by default in the self-attention module in Swin Blocks to help encode the relative position of pixels. Sinusoidal position encoding (SPE) [66] admits translation invariance and we add it to the feature maps in the downsampling and upsampling layers. Through the step-by-step downsampling operation, multi-scale content and style representations at feature levels can be learned by content and style encoders, adaptively fitting different feature transformations.

2) Generator: Fig. 2 illustrates our generator architecture. We take a noise map that is sampled from a gaussian distribution as the latent input $\boldsymbol{Z_0}$. The generator is designed to consistently match the different scales of content and style features. The generation process aims to decode high-level

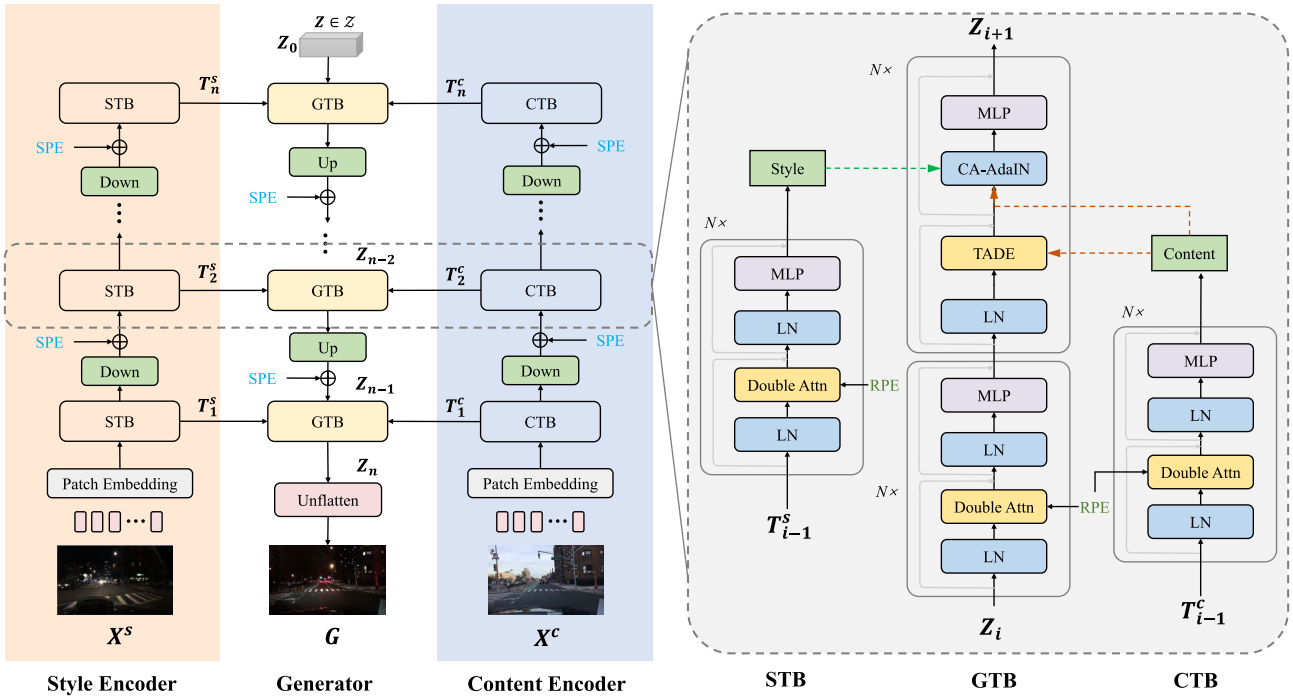

Fig. 2. (a) The proposed SwinIT architecture. Content and style encoders extract the multi-scale feature representations of content image $\boldsymbol{X^c}$ and style image $\boldsymbol{X^s}$ respectively. The content and style features at the same scale are fused to generate an image $\boldsymbol{G}$ in an end-to-end training. The gray box on the right describes the details of components (e.g., **S**tyle, **G**enerator, and **C**ontent **T**ransformer **B**locks) at the same level. Submodules of our network are shown in Fig. 3.

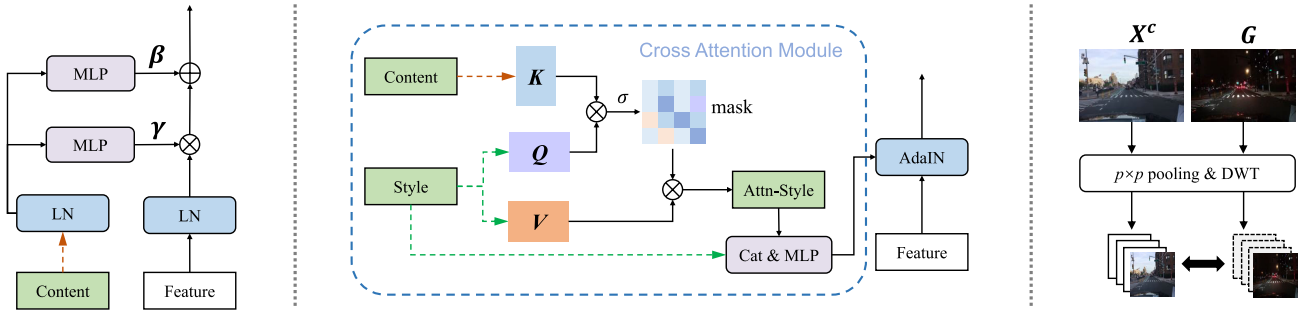

Fig. 3. Submodules of our framework. (a) is a Transformer adaptive denormalization (TADE) module for content feature transformation. It uses modulation parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ learned by the content feature to perform element-wise denormalization. (b) is a cross-attention adaptive instance normalization (CA-AdaIN) module for effective style injection. It performs cross-attention mechanisms to achieve precise style adaptive normalization. (c) is the proposed WTM loss, $\boldsymbol{X^c}$ and $\boldsymbol{G}$ denote the content and generated image, respectively.

latent codes into low-level representations. It receives a latent variable $\boldsymbol{Z_0} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ as the input and gradually upsamples the feature maps through a cascade of Transformer blocks, finally getting the generated image $\boldsymbol{G}$.

The feature maps from corresponding layers in content or style flow and generator are fused by a TADE block and a CA-AdaIN module in end-to-end training. As is shown in Fig. 3, TADE uses modulation parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ learned by the content feature to perform element-wise denormalization. CA-AdaIN performs a cross-attention mechanism to achieve precise style adaptive normalization.

3) Multi-Scale Discriminator: We exploit two patch-based discriminators [15], [27] with an identical architecture to discriminate images at different scales. A multi-scale patch-based strategy allows discriminators to capture coarse-to-fine information about the image in different receptive fields. This helps the generator produce high-frequency details of images in translation tasks. In addition, the discriminator is also used as a feature extractor for the generator to optimize the feature matching loss.

B. Hierarchical Content and Style Injection

We propose a new feature transformation scheme that fuses content and style information adaptively. Let $\boldsymbol{Z_i} \in \{\boldsymbol{Z_0}, \boldsymbol{Z_1}, \ldots, \boldsymbol{Z_n}\}$ represent the output feature maps of the i-th residual block in the generator, where n denotes the total number of residual blocks and $\boldsymbol{Z_0}$ is a latent input for the generator, which is sampled from a Gaussian distribution. Let $\boldsymbol{T_i^c} \in \{\boldsymbol{T_0^c}, \boldsymbol{T_1^c}, \ldots, \boldsymbol{T_n^c}\}$ be the hierarchical content representations extracted by the content encoder, and $\boldsymbol{T_i^s} \in \{\boldsymbol{T_0^s}, \boldsymbol{T_1^s}, \ldots, \boldsymbol{T_n^s}\}$ be the hierarchical style representations extracted by the style encoder, as shown in Fig. 2.

1) Transformer Adaptive Denormalization (TADE): Common normalization layers (BN, IN) are essential components that can stabilize the input distribution and accelerate the degree of training convergence. But for the generation network of the image-to-image translation task, these unconditional normalization layers will tend to wash out semantic information when applied to uniform or smooth semantics. Therefore, we need to perform conditional normalization by learning the affine transformation parameters of the input distribution to control the global spatial variation. Different from SPADE [1] that takes a semantic mask as the input of each resblock, we turn the input into a multi-scale feature representation $\boldsymbol{T_i^c}$ of the content image $\boldsymbol{X^c}$ by a series of Transformer blocks. Also, our Transformer-based model is fed with patchwise images, which means that the same position of different images in the batch may have different distributions. Hence, layer normalization (LN) is more suitable than batch normalization [1] to learn spatial semantics. In this way, we take full advantage of the semantic information captured by the content encoder. Formally, we define C_i, H_i, W_i as the number of channels, height size, and width size of feature maps on the i-th layer. As shown in Fig. 3(a), we first apply layer normalization to normalize the content $\boldsymbol{T_i^c}$ and the feature $\boldsymbol{Z_{n-i}}$. Then we modulate the normalized feature by using the parameters scale $\boldsymbol{\gamma_i}$ and bias $\boldsymbol{\beta_i}$ learned from the content $\boldsymbol{T_i^c}$. The denormalized process is formulated as:

$$\hat{\boldsymbol{Z}}_{\boldsymbol{n-i}}^{\boldsymbol{c,h,w}} = \boldsymbol{\gamma_i^{c,h,w}} \cdot \mathrm{LN}(\boldsymbol{Z_{n-i}^{c,h,w}}) + \boldsymbol{\beta_i^{c,h,w}}, \tag{2}$$

where $c \in C_i, h \in H_i, w \in W_i$, the denormalization operation is element-wise. Compared to previous normalization works [1], [26], TADE can better preserve semantic content information with the hierarchical Transformer-based structure.

2) Cross-Attention Adaptive Instance Normalization (CA-AdaIN): Unpaired I2I translation often suffers from the semantic mismatch between source and reference examples. Previous style normalization strategies [26], [27] can merely generate images with overall similarities to the reference. To address this issue, we introduce CA-AdaIN to better fuse style and content representations. Different from AdaIN which generalizes feature maps channel-wise, we enable the style encoder to learn the multi-scale feature-level style representation $\boldsymbol{T_i^s}$ of the style image $\boldsymbol{X^s}$ more effectively. Furthermore, the content representation $\boldsymbol{T_i^c}$ is used to query and calculate the correlation between the semantic information and style at the patch level. The weighted style representation with a self-attention mechanism is sent to generalize the denormalized feature map $\hat{\boldsymbol{Z}}^{\boldsymbol{n-i}}$ by AdaIN. The process is formulated as:

$$\begin{aligned} \boldsymbol{T'^s_i} &= \text{W-MSA}(\boldsymbol{W_Q^c} \cdot \boldsymbol{T_i^c}, \boldsymbol{W_K^s} \cdot \boldsymbol{T_i^s}, \boldsymbol{W_V^s} \cdot \boldsymbol{T_i^s}), \\ \hat{\boldsymbol{T}}_{\boldsymbol{i}}^{\boldsymbol{s}} &= \mathrm{MLP}(\mathrm{Concat}(\boldsymbol{T'^s_i}, \boldsymbol{T_i^s})), \end{aligned} \tag{3}$$

where $\boldsymbol{W_Q^c}, \boldsymbol{W_K^c}, \boldsymbol{W_V^c} \in \mathcal{R}^{(H_i W_i)^2 \times C_i}$ are the query of the content, key, and value of the style matrices, and W-MSA denotes window-based multi-head self-attention.

We use $\hat{\boldsymbol{Z}}_{\boldsymbol{i}}$ to represent the feature map after the i-th TADE transformation and compute the affine parameters from the style feature $\hat{\boldsymbol{T}}_{\boldsymbol{i}}^{\boldsymbol{s}}$ after the Cross-Attention mechanism.

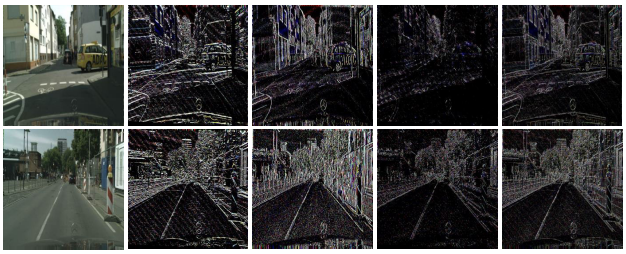

Fig. 4. An illustration of wavelet transformation. $\boldsymbol{H}$ denotes high-pass filtering, while $\boldsymbol{L}$ represents low-pass filtering.

Compared with AdaIN [26], CA-AdaIN experiences more flexible influence with a conditional semantic information manner at a multi-scale feature level, thus it can adaptively transfer styles more precisely.

$$\text{CA-AdaIN}(\hat{\boldsymbol{Z}}_{\boldsymbol{n-i}}, \hat{\boldsymbol{T}}_{\boldsymbol{i}}^{\boldsymbol{s}}) = \sigma(\hat{\boldsymbol{T}}_{\boldsymbol{i}}^{\boldsymbol{s}})\left(\frac{\hat{\boldsymbol{Z}}_{\boldsymbol{n-i}} - \boldsymbol{\mu}(\hat{\boldsymbol{Z}}_{\boldsymbol{n-i}})}{\sigma(\hat{\boldsymbol{Z}}_{\boldsymbol{n-i}})}\right) + \mu(\hat{\boldsymbol{T}}_{\boldsymbol{i}}^{\boldsymbol{s}}), \tag{4}$$

where μ and σ are the mean and standard deviation, respectively. Exploiting cross-attention adaptive instance normalization (CA-AdaIN), coarse-to-fine style features at different layers can be fused adaptively with the corresponding semantic structure features learned by Transformer adaptive denormalization (TADE), allowing our framework to be trained end-to-end and versatile for different tasks.

C. Objective Function

1) Wavelet Transformation: In the field of computer vision, wavelet transform is an effective tool for decomposing images into multi-frequency domain descriptions. Consequently, we adopt a two-dimensional discrete wavelet, the Haar wavelet, which can perform low-pass ($L^T = \frac{1}{\sqrt{2}}[1, 1]$) and high-pass ($H^T = \frac{1}{\sqrt{2}}[-1, 1]$) filtering from horizontal and vertical directions. Haar wavelet transformation includes four filters, {***LL, LH, HL, HH***}, which emphasize content information, horizontal, vertical, and diagonal edges or gradients, respectively. Among them, ***LH***, ***HL***, and ***HH*** contain the information that represents rich details (see Fig. 4), which helps our model recover details and suppress unpleasant artifacts with supervised learning from real images.

2) Losses: We use standard and succinct losses in our objective function. We apply the hinge-based adversarial loss to distinguish between the real and fake images for the generator and the multi-scale discriminators. Meanwhile, perceptual loss and feature matching loss are applied to the generator.

However, the above losses only catch the feature-level style or content reconstruction for the generated image, but cannot protect explicit structural scene details (e.g., high-frequency edges and textures) and only preserve rough style features after the optimization. The traditional optimization objectives [1], [27], [28] are either coarse-grained or sensitive to small perturbations. To solve this problem, we apply the wavelet transformation at the image level, which decomposes images into different frequency bands to capture rich details under the supervision of a real image. Specifically, as shown in Fig. 3(c),

we first smooth the original image with a $p \times p$ pooling layer to prevent the filter from being too sensitive to small perturbations of the input image. Then we adopt the above-mentioned wavelet transformation to extract multi-frequency domain features from the pooled image. Finally, we apply a wavelet transformation matching (WTM) loss to the generator, which is defined as:

$$\mathcal{L}_{\mathbf{W}}(\boldsymbol{X^c}, \boldsymbol{X^s}) = ||\Psi(Gen(\boldsymbol{X^c}, \boldsymbol{X^s})) - \Psi(\boldsymbol{X^{c/s}})||_1, \quad (5)$$

where Ψ is the operation of the wavelet transform with smoothing. The reference image is $\boldsymbol{X^s}$ in the semantic image synthesis task and $\boldsymbol{X^c}$ in other tasks. *Gen* is the translation model (in Fig. 2) that can synthesize a fake image from the input content and style images. We apply MAE loss, which has better robustness to outliers. The overall loss function is the weighted summation of the losses:

$$\begin{aligned}\mathcal{L}_G &= -\mathbb{E}[log\, D(\boldsymbol{G})] + \lambda_P \mathcal{L}_P(\boldsymbol{G}, \boldsymbol{X^c}) \\ &\quad + \lambda_{FM}\mathcal{L}_{FM}(\boldsymbol{G}, \boldsymbol{X^s}) + \lambda_W \mathcal{L}_W(\boldsymbol{G}, \boldsymbol{X^c}), \\ \mathcal{L}_D &= -\mathbb{E}[min(0, -1 + D(\boldsymbol{X^s}))] \\ &\quad - \mathbb{E}[min(0, -1 - D(\boldsymbol{G}))], \end{aligned} \quad (6)$$

where $\boldsymbol{G} = Gen(\boldsymbol{Z_0}, \boldsymbol{X^c}, \boldsymbol{X^s})$ denotes the generated image, λ_P, λ_{FM} and λ_W are the corresponding weights, respectively. $\mathcal{L}_P$ is the perceptual loss that minimizes the style difference between $\boldsymbol{G}$ and $\boldsymbol{X^c}$ in multi-scale feature levels extracted by pre-trained VGG-19 networks. For the feature matching loss $\mathcal{L}_{FM}$, the intermediate features at different layers of $\boldsymbol{G}$, $\boldsymbol{X^s}$ are matched by the off-the-shelf discriminators. The objective functions make our framework stable and easy to train without complex cycle consistency constraints.

IV. Experiments

A. Implementation Details

In the experiment, we optimize the model using Adam [72] and set $\beta_1 = 0$ and $\beta_2 = 0.9$. The weights of networks use a xavier distribution in the initialization, and the learning rates for the generator and discriminator are set to 0.0001 and 0.0004, respectively. We adopt LeakyReLU activations with a slope of 0.2 in the discriminator. We apply spectral normalization [73] to layers in the networks to ensure Lipschitz continuity. For the perceptual loss, we follow the settings in TSIT [27]. The experiments are conducted on NVIDIA RTX 2080ti GPUs.

B. Datasets

We explore eight datasets in diverse scenarios to verify the effectiveness of our method. For all experiments, the resolution of the input images is scaled to 256×256 and the generated images are 256×256 as well. For arbitrary style transfer, we use Yosemite Summer→Winter dataset [13], Day→Night in BDD100K [74] dataset, and Photo→Art dataset [13]. For semantic image synthesis, we select several challenging datasets, including CMP Facades [75], Map→Aerial photo dataset [15], Cityscapes [76] and ADE20K [77]. For multi-modal image synthesis, we further classify BDD100K into different time and weather conditions and perform controllable time and weather translation.

1) Yosemite Summer→Winter: We carry out the season-style transfer task for unpaired I2I translation. Similar to the experimental setup of CycleGAN [13], the dataset consists of 1,231 summer images and 962 winter images for training.

2) Day→Night in BDD100K Dataset: We perform time translation with 12,454 daytime images and 22,884 nighttime images. The domain labels of *day* and *night* come from the BDD100K dataset. All the images are captured on the road or in street scenes with an original resolution of $1,280 \times 720$.

3) Photo→Art: The dataset contains 6,287 photos of real scenes and 2,559 artistic paintings by several artists (e.g., Cezanne, Monet, van Gogh, and Ukiyo-e) for training. Different from CycleGAN, we combine the paintings of all the artists to show the robustness of our method.

4) Map→Aerial Photo: The task is evaluated on data scraped from the Google Maps dataset, which consists of 1,098 images in both the training set and testing set. Meanwhile, each image includes paired aerial photos and maps.

5) Facades label→Photo: Facades is a dataset of building appearance images assembled at the Machine Perception Center, comprising 606 rectified images from various sources that have been manually annotated. The images are from cities and different architectural styles from all over the world. In our experiments, we take 400 images for training and 106 images for testing.

6) Sunny→Different Weathers in the BDD100K Dataset: Similar to the collection of Day→Night, we filter BDD100K into different domains and perform multi-modal weather translation in real-word scenes. The training set consists of 10,000 sunny images and 10,000 different weather images (e.g., night, cloudy, rainy, and snowy).

7) Cityscapes Dataset: The Cityscapes dataset mainly includes images of urban street scenes, which consist of 50 different cities in Germany. It contains 2,975 pairs of images for training and 500 images for evaluation, respectively. The initial resolution of images is $2,048 \times 1.024$. The dataset provides high-quality, pixel-level annotations for 30 classes.

8) ADE20K Dataset: ADE20K dataset is a challenging scene dataset with indoor and outdoor scenes for various segmentation tasks. It consists of 20,210 images and 2,000 images for training and validation, respectively, with fine annotations for 150 semantic classes.

C. Metrics

We employ the following standard metrics to quantitatively evaluate the synthesized images. Our setup is consistent with prior works [13], [27], [68] for fair comparison.

We use the Frechet Inception Distance (FID) [78] for arbitrary style transfer and semantic image synthesis. FID considers the similarity between the distribution of the real image and the generated image, lower is better. This measurement respectively extracts the 2048-dim feature vectors of images from the pre-trained Inception-V3 [79] network. Furthermore, we use the Inception Scores (IS) [80] and Learned Perceptual Image Patch Similarity (LPIPS) [81] to evaluate the diversity and clarity of generated images for arbitrary style transfer, higher is better. LPIPS uses pre-trained AlexNet [82] that

TABLE I

THE FID, IS, AND LPIPS SCORES OF OUR METHOD COMPARED TO STATE-OF-THE-ART METHODS IN ARBITRARY STYLE TRANSFER TASKS. A LOWER FID AND A HIGHER IS OR LPIPS INDICATE BETTER PERFORMANCE

Method	*Summer→Winter*			*Day→Night*			*Photo→Art*		
	FID (↓)	IS (↑)	LPIPS (↑)	FID (↓)	IS (↑)	LPIPS (↑)	FID (↓)	IS (↑)	LPIPS (↑)
CycleGAN [13]	77.76	3.00	0.351	28.86	2.29	0.516	124.4	3.27	0.311
TSIT [27]	80.16	2.69	0.301	30.59	2.16	0.492	125.1	3.15	0.438
DMIT [54]	87.97	2.88	0.277	83.90	2.16	0.428	166.9	3.47	0.392
MUNIT [16]	118.2	2.54	0.419	110.0	2.19	0.469	167.3	3.28	0.371
SAVI2I [67]	86.02	2.85	0.274	126.2	2.52	0.452	144.9	3.50	0.449
QS-Attn [68]	148.3	2.95	0.302	47.57	**2.60**	0.536	123.5	3.25	0.437
AdaAttN [49]	112.8	2.66	0.315	84.71	2.41	0.510	114.1	2.82	0.477
SRIT [69]	70.50	2.99	0.422	65.86	2.33	0.391	98.72	3.03	0.481
CAST [70]	82.65	2.85	0.371	70.74	2.26	0.377	102.9	3.11	0.480
Ours (SwinIT)	**64.60**	**3.16**	**0.427**	**19.36**	2.54	**0.552**	**81.37**	**3.53**	**0.489**

TABLE II

THE MIOU, PIXEL/CLASS ACCURACY, AND FID SCORES OF OUR METHOD COMPARED TO STATE-OF-THE-ART METHODS IN SEMANTIC IMAGE SYNTHESIS TASKS. A HIGHER MIOU, A HIGHER PIXEL/CLASS ACCURACY, AND A LOWER FID INDICATE BETTER PERFORMANCE

Method	*Cityscapes*				*ADE20K*			
	mIoU (↑)	Per-pixel acc. (↑)	Per-class acc. (↑)	FID (↓)	mIoU (↑)	Per-pixel acc. (↑)	Per-class acc. (↑)	FID (↓)
CycleGAN [13]	11.7	50.8	18.4	76.31	14.2	45.9	19.8	96.68
SPADE [1]	20.8	80.3	24.6	71.86	38.5	79.9	42.8	**33.92**
Pix2pixHD [51]	20.5	78.5	24.1	95.05	20.3	69.2	25.2	81.87
CRN [71]	18.9	68.6	20.1	104.7	22.4	68.8	22.7	73.32
TSIT [27]	20.6	79.3	24.4	74.31	38.6	80.8	44.1	35.21
QS-Attn [68]	16.4	66.2	24.0	75.66	10.7	55.5	13.1	107.3
Ours (SwinIT)	**21.8**	**80.8**	**26.7**	**66.10**	**39.1**	**80.8**	**44.4**	37.13

highly agrees with humans' perception. We compute the LPIPS metric as the average of 10 images translated from the same source image.

For semantic image synthesis, we adopt segmentation accuracy (mean Intersection-over-Union (mIoU), per-pixel accuracy, and per-class accuracy). It utilizes an off-the-shelf semantic segmentation algorithm to predict a label map of the generated image. We take the pre-trained FCN [83] for Cityscapes [76] and UperNet101 [84] for ADE20K [77]. Moreover, we adopt the Peak Signal Noise Ratio (PSNR) to measure the similarity between a synthesized image and the corresponding target image. Higher PSNR values indicate better quality.

D. Baselines

We make both qualitative and quantitative comparisons between our method and several state-of-the-art baselines. We conduct experiments on nine baseline approaches for style transfer tasks in unsupervised settings, including CycleGAN [13], TSIT [27], QS-Attn [68], SAVI2I [67], DMIT [54], MUNIT [16], AdaAttN [49], SRIT [69], and CAST [70]. MUNIT, DMIT, and SRIT have the ability to capture the multimodal nature of images while maintaining quality. For the baselines [16], [54], [67] that cannot complete multitasking, we replace them with some recent works in the semantic synthesis task. We compare CycleGAN, TSIT, QS-Attn, SPADE [1], CRN [71], and Pix2pixHD [51] in the supervised settings. We import the pre-trained models if possible. Otherwise, we train their models from scratch and try our best to tune them using officially released source codes.

E. User Preference Study

We perform a user study to evaluate the visual fidelity of the generated images. Following the evaluation protocol of SPADE [1], we use the Amazon Mechanical Turk (AMT) to compare our method against three baselines (CycleGAN [51], TSIT [27], and QS-Attn [68]). The results are shown in Table III. For each dataset, five subjects were shown with 40 different samples generated using our model or a baseline model. They are asked to choose the output image that looks more realistic when given the corresponding segmentation mask.

Compared with CycleGAN [13] and QS-Attn [68], the participants mostly preferred our results. This gap can be attributed to the limitations of cycle consistency and unrobust contrastive learning strategies, respectively. As compared to TSIT [27], the preferences of the participants differed relatively little between both methods for most datasets. Overall, we can see that our model compares favorably against the three state-of-the-art baselines on all the datasets, further validating that our method achieves higher image generation quality.

F. Quantitative Evaluation

The quantitative evaluation results compared with several leading methods are shown in Table I, II, V and IV.

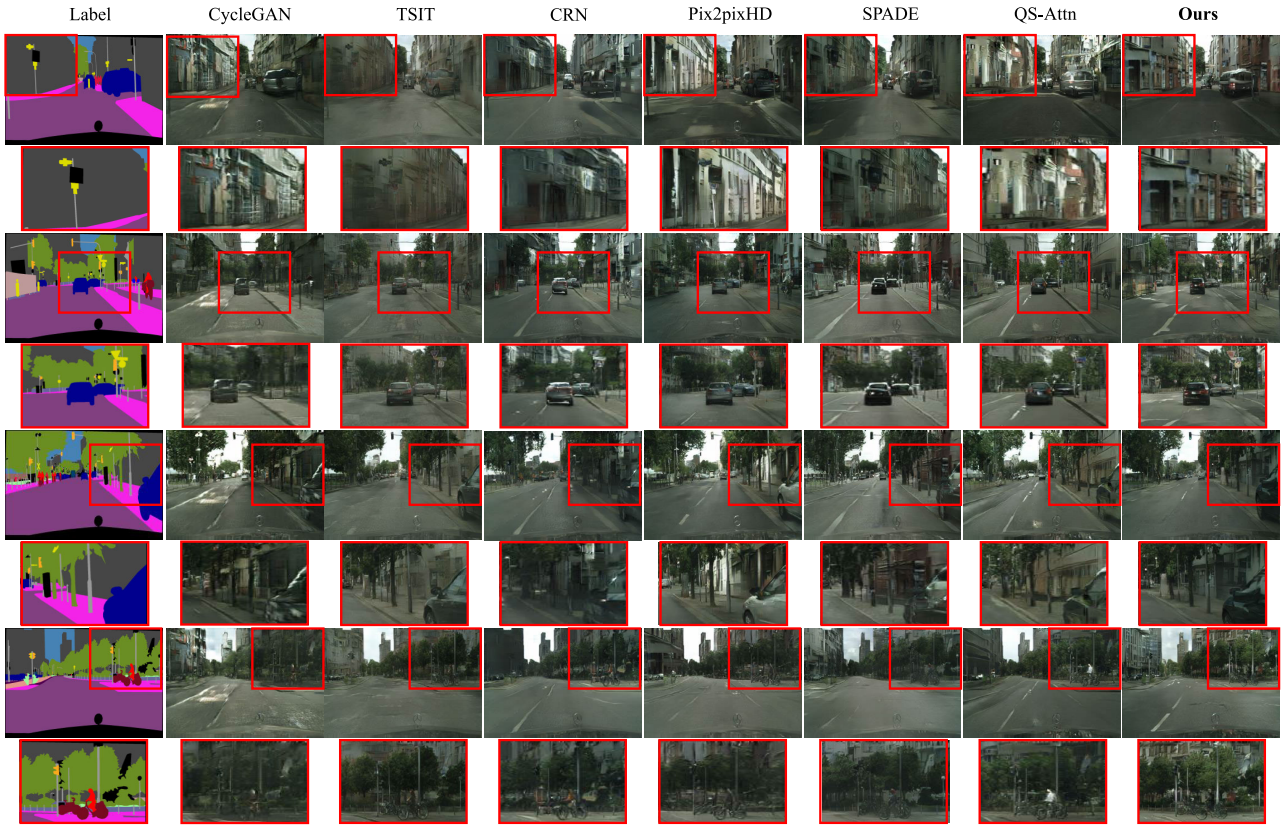

Fig. 5. Visual results on Cityscapes compared with baselines. The leftmost column contains the input source images. In the remaining columns, from left to right, are the translated results of our model and baselines. The local regions in the red frame are enlarged.

TABLE III

USER STUDY. THE NUMBERS INDICATE THE PERCENTAGE OF USERS WHO FAVOR THE RESULTS OF THE PROPOSED METHOD OVER THOSE OF THE COMPETING METHOD. OUR METHOD ACHIEVES BETTER PERFORMANCE

AMT (↑)	*S2W*	*D2N*	*P2A*	*S2D*	*Cityscapes*	*ADE20K*	*Average*
Ours vs. CycleGAN [13]	68.26	64.33	57.18	-	78.42	76.59	68.96
Ours vs. TSIT [27]	70.04	59.86	53.41	58.10	62.85	54.11	59.73
Ours vs. QS-Attn [68]	83.66	87.41	71.94	-	69.05	84.91	79.39

Our method achieves better performance than other baselines in several datasets. In arbitrary style transfer tasks, we test the quality and diversity of the generated images through FID, IS, and LPIPS. Our method realizes considerable performance gains. It should be noticed that a 19% drop for summer → winter, 37% drop for day → night, and 35% drop for photo → art represent the superior performance of our Transformer-based GAN architecture than CNN-based TSIT [27]. Moreover, different from the observation in IPT [60] that Transformer-based models rely on a large amount of training data, SwinIT achieves better results than CNN-based models using the same training data without any pretraining process, even when the dataset is small (i.e., Yosemite summer → winter).

In semantic image synthesis tasks, our approach achieves state-of-the-art semantic segmentation performance through FCN and FID scores on Cityscapes. On the ADE20k dataset, the proposed method can achieve comparable performance with the specialized leading methods [1], [27], [51] on semantic segmentation metrics. Note that in the rest of the generic baseline methods [13], [68], our method outperforms them by a large margin, suggesting its strong robustness to fit the nature of different I2I translation tasks. In the classic and paired image datasets of Maps and Facades, SwinIT also achieves the best performance under the FID and PSNR metrics.

In the multi-modal image synthesis task on BDD100K, we translate the images of sunny weather to different times and weathers (i.e., night, snowy, cloudy, rainy) with only a single model. Table IV demonstrates that our model achieves better FID scores, especially for the sunny→night mapping.

G. Qualitative Evaluation

We compare our method with several state-of-the-art baselines to verify the performance of arbitrary style transfer tasks in diverse scenarios (e.g., natural images, real-world scenes, and artistic paintings). The results are shown in Fig. 6. The baselines sometimes transfer an incomplete style to the whole image, which produces visually unrealistic results. Among them, SAVI2I cannot perform any style change at all, while TSIT produces an unreasonable winter style. MUINT tends to impose a uniform color on the image. In contrast, our

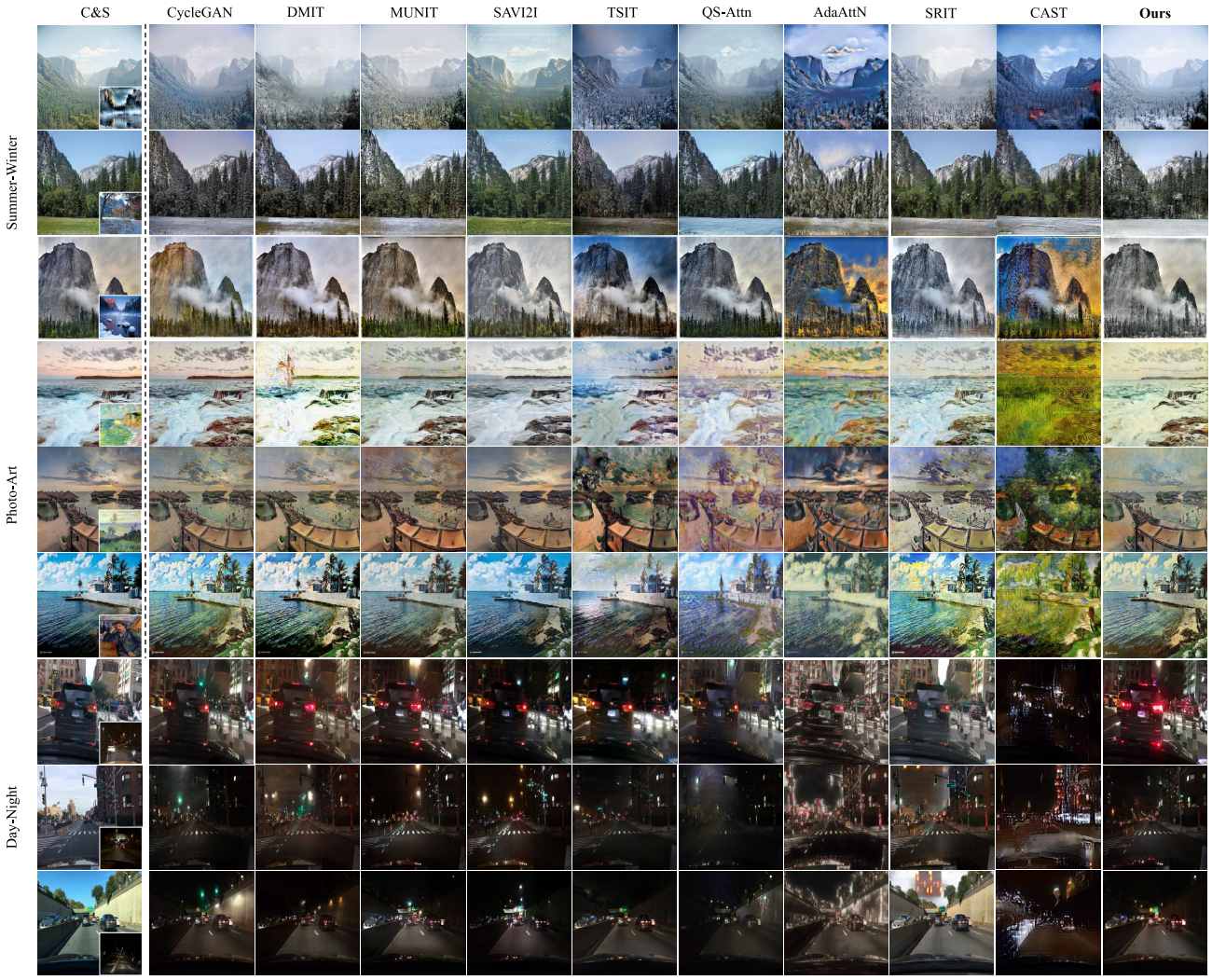

Fig. 6. Arbitrary style transfer (Yosemite summer → winter, photo → art, and BDD100K day → night) results compared to baselines. C&S denotes content and style.

method generates more pleasing results that depend on the target domain style. In the results of the day → night on the BDD100K dataset, the generated results of QS-Attn and DMIT are mostly vague, meaning that the content representations are not well preserved. TSIT, CycleGAN, AdaAttN, and MUNIT fail to transfer the style of key objects (e.g., cars, road signs) but produce some unreasonable highlighted artifacts. In comparison, our method generates more photorealistic and clearer samples. In the task of photo → art, SwinIT transfers the styles well while effectively keeping the content structure. However, the baselines such as CAST bring some blocking artifacts. MUNIT, SRIT, and SAVI2I lead to style degeneration.

We also perform multi-modal image translation tasks on the BDD100K dataset with a single model. As shown in Fig. 8, the model translates the images from *sunny* to different weathers or times (e.g., *night, cloudy, rainy, and snowy*). DMIT and SRIT lead to style degeneration on sunny→night mapping, while MUINT generates the image with the wrong style (darkness) on sunny→cloudy mapping. Our method can successfully achieve several mappings (i.e., sunny → night, sunny → cloudy, and sunny → rainy) and produce photorealistic quality. It is noticed that the sunny→snowy mapping sometimes cannot generate snow-like styles well. We attribute it to the tiny snow-style area and the fact that the style of the snow in the road traffic scene is vandalized and lacks regularity.

TABLE IV

THE FID SCORES OF OUR METHOD COMPARED TO STATE-OF-THE-ART METHODS IN MULTIMODAL IMAGE SYNTHESIS TASKS. A LOWER FID INDICATES BETTER PERFORMANCE. WE TAKE SUNNY AS OUR SOURCE DOMAIN (CONTENT)

Method	*Night*	*Cloudy*	*Rainy*	*Snowy*	*Average*
MUNIT [16]	100.3	30.76	62.26	87.55	70.22
DMIT [54]	89.59	49.75	50.30	69.48	64.78
SRIT [69]	75.28	**21.18**	55.43	74.96	56.71
Ours (SwinIT)	**32.76**	29.41	**47.15**	**61.20**	**42.63**

In semantic image synthesis tasks, the results of the Cityscapes dataset are shown in Fig. 5. Our method can synthesize reasonable street scenes, including fine-grained local objects (e.g., cars, road signs, and pedestrians). TSIT generates high-frequency texture features but lacks the necessary semantics. CRN, CycleGAN, and QS-Attn generate

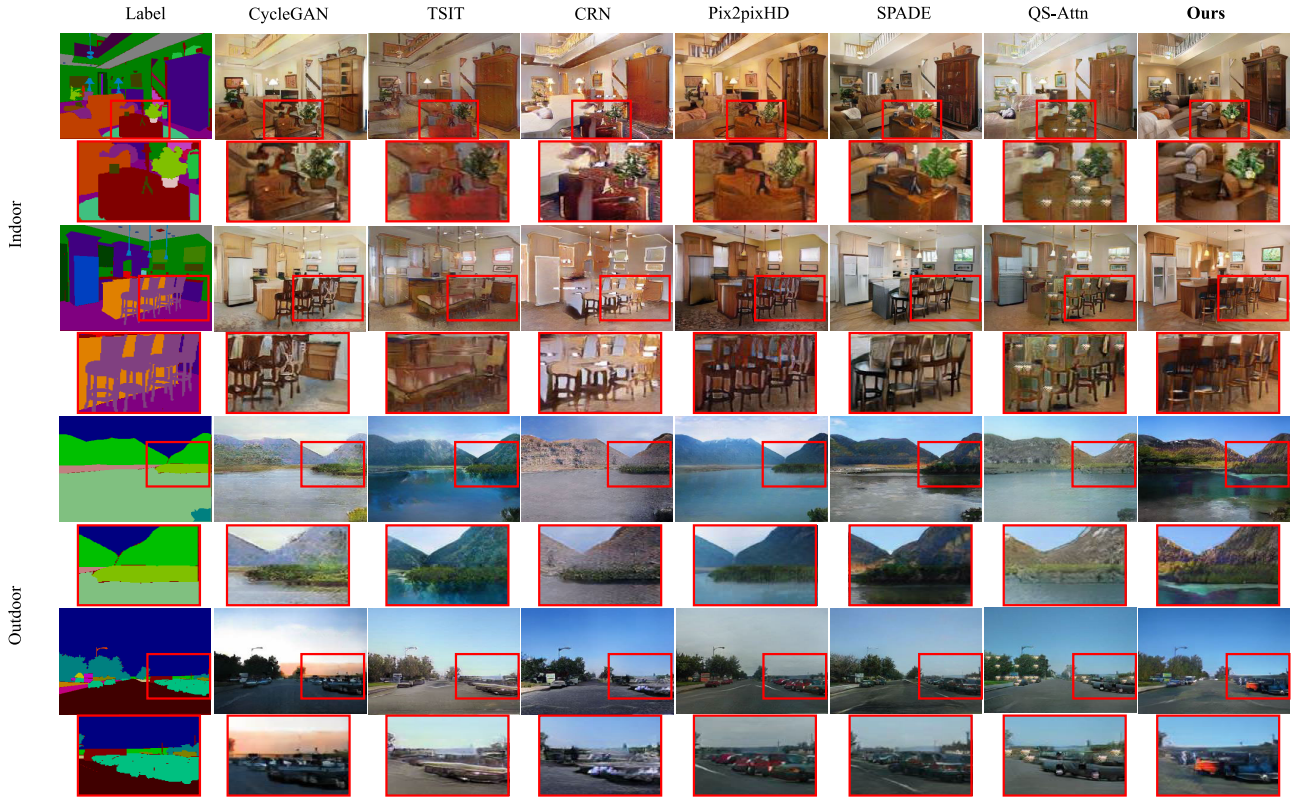

Fig. 7. Visual results on ADE20K compared with baselines. The leftmost column contains the input source images. In the remaining columns, from left to right, are the translated results of our model and baselines. The local regions in the red frame are enlarged.

TABLE V

THE FID SCORES AND PSNR OF OUR METHOD COMPARED TO STATE-OF-THE-ART METHODS IN PAIRED IMAGE SYNTHESIS TASKS. A HIGHER PSNR AND A LOWER FID INDICATE BETTER PERFORMANCE

Method	*Map→Photo*		*Label→Photo*	
	FID (↓)	PSNR (↑)	FID (↓)	PSNR (↑)
CycleGAN [13]	35.65	27.27	26.2	27.63
SPADE [1]	41.93	27.58	93.51	28.29
Pix2pixHD [51]	38.78	27.62	98.76	28.45
CRN [71]	57.11	27.03	107.4	28.02
TSIT [27]	72.69	28.12	91.90	28.64
QS-Attn [68]	78.80	27.55	110.5	27.97
Ours (SwinIT)	**34.54**	**28.29**	**90.03**	**28.85**

unreasonable structural information, such as buildings. SPADE and Pix2pixHD can fit the background (roads and buildings) well but produce distortions on local objects in some cases. In indoor scenes of ADE20k, TSIT and Pix2pixHD generate unreasonable distortion in the semantic synthesis of some objects (e.g., chairs and tables) with complex shapes. Our method generates better results due to the supervision of wavelet-based edges. For outdoor scene images, CRN and QS-Attn produce unpleasing generation qualities, such as grass and water. Our generated images have higher color saturation and better diversity. Compared to other baselines, our method is versatile for diverse scenarios and shows better consistency with ground truths.

Similar comparison results are also obtained on the CMP Facades dataset and the Google Maps dataset. More visual comparison results are in the supplementary materials.

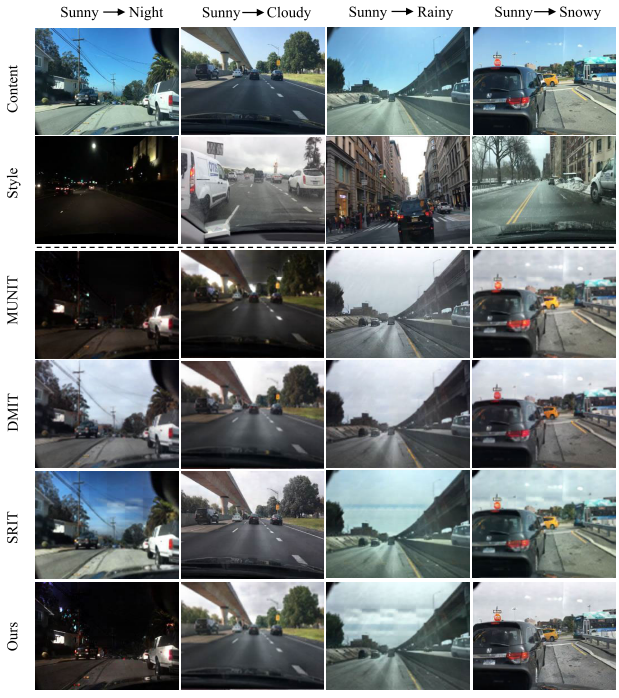

Fig. 8. BDD100K multi-modal image synthesis with different time and weather translation results by a single model.

H. Ablation Studies

1) Effectiveness of Specific Modules: We present additional ablation study results for individual components of the proposed method on several datasets. Table VI provides quantitative evaluation results using the FCN scores, FID,

TABLE VI

QUANTITATIVE RESULTS FOR ABLATION STUDIES ON SEMANTIC IMAGE SYNTHESIS AND ARBITRARY STYLE TRANSFER TASKS. IN CONFIGURATION, MODEL MEANS OPTIONAL KEY MODULES IN OUR FRAMEWORK AND ADOPTING THEM TO TRANSFORM INTERMEDIATE FEATURES.WTM DENOTES SELECTING DIFFERENT POOLING OPERATIONS (E.G., NO DOWNSAMPLING, MAX POOLING, AND AVERAGE POOLING) ON IMAGES BEFORE APPLYING WAVELET TRANSFORMATION MATCHING LOSS.MODEL G IS THE SETTINGS CORRESPONDING TO THE LAST ROW IN TABLE II

Method	Configuration							*Cityscapes*			*ADE20K*			*Yosemite*	
	Model				WTM			FID (↓)	Acc (↑)	mIoU (↑)	FID (↓)	Acc (↑)	mIoU (↑)	FID (↓)	IS (↑)
	SPADE	TADE	AdaIN	CA-AdaIN	No DS.	MP.	AP.								
A	✓		✓					108.4	67.9	15.1	94.90	63.7	16.6	102.9	2.03
B		✓	✓					91.17	73.2	15.9	75.38	68.9	19.3	87.65	2.46
C	✓			✓				76.09	78.7	19.6	79.07	69.4	21.1	89.96	2.45
D		✓		✓				71.88	77.3	18.5	64.76	71.1	22.5	82.18	2.73
E		✓		✓	✓			70.61	79.9	21.1	53.71	76.3	33.4	69.01	3.07
F		✓		✓		✓		**63.35**	79.2	20.6	40.96	80.6	38.7	67.95	3.11
G		✓		✓			✓	66.10	**80.8**	**21.8**	**37.13**	**80.8**	**39.1**	**64.60**	**3.16**

Fig. 9. Validation of the ineffectiveness of task-specific methods with cycle consistency settings.

TABLE VII

COMPARISON OF THE NETWORK PARAMETERS, MACS, AND FID WITH TSIT AND BASELINES. WE USE STYLE TRANSFER TASK FOR ABLATIONS

Methods	Blocks	Settings	#Params	MACs	FID (↓)
TSIT [27]	-	-	155.53M	37.99G	30.59
A.	6	AdaIN	164.16M	18.60G	54.18
B.	6	CA-AdaIN	176.75M	20.00G	**19.36**
C.	1	B.	13.165M	5.186G	188.4
D.	2	B.	13.650M	8.180G	162.9
E.	3	B.	15.578M	11.15G	111.2
F.	4	B.	23.267M	14.11G	89.55
G.	5	B.	53.982M	17.05G	41.41

and IS to analyze the contribution of key modules of the framework, including feature transformation and optimization objectives. In arbitrary style transfer task, Transformer adaptive denormalization (TADE) and Cross-Attention AdaIN (CA-AdaIN) lead to performance gains from models A to D. We also find that applying wavelet transformation matching loss $\mathcal{L}_W$ can improve the results when comparing the models D and E. Moreover, Models F and G outperform Model C, reflecting the effectiveness of smoothing original images with a pooling operation (e.g., max pooling and average pooling). Furthermore, model G achieves the best scores among all the previous models. We attribute them to the fact that average pooling can preserve the global information of the image, while max pooling focuses more on textures or edges. In the general indoor and outdoor scenes of semantic image synthesis, Table VI demonstrates that both TADE and CA-AdaIN lead to quality and diversity gains from models A to D. However, it seems that SPADE is more suitable than TADE. In the special road traffic scene of the semantic image synthesis task, the introduction of the referenced mask image can better generate the texture details or semantic information of the image than abstract features. Directly inputting the semantic mask into the intermediate layer helps the model learn low-level, aligned, detailed texture information. This is due to the particularity of the scene, and the SPADE [1] only validates the datasets of semantic image synthesis, while our method is more versatile for diverse tasks.

From the qualitative evaluation perspective, we also provide the visualization results. We perform semantic image synthesis, arbitrary style transfer, and multi-modal image synthesis tasks on several examples to verify the effectiveness of different modules. More visual and quantitative comparative results and analyses are in the supplementary materials.

2) Parameters and Throughput: We compare the number of model parameters and MACs with TSIT for 256 × 256 translations, as shown in Table VII. Although our approach has a larger model size, it achieves lower MACs than TSIT, which means the method achieves competitive generation quality with less theoretical computational cost. Moreover, the Cross-Attention mechanism increases the number of parameters by less than 8% and achieves a large performance gain. By a quantitative analysis of the number of the Transformer blocks, with the sharp drop in the number of parameters, the performance of the model is also significantly affected. It illustrates the effectiveness of capturing and fusing content and style representations in a coarse-to-fine manner in our framework.

3) Evaluations of Hyper Parameters: For the ablation study of the architecture, we train our SwinIT on different scales on three style transfer tasks.

a) Selection of patch number of the input image: We first exhibit the influence of the number of patches (P) on the quality of the generated images in Fig. 10(a). When fixing the rest of the parameters, it can be observed that the FID and IS gains become larger when the patch size is smaller. Therefore,

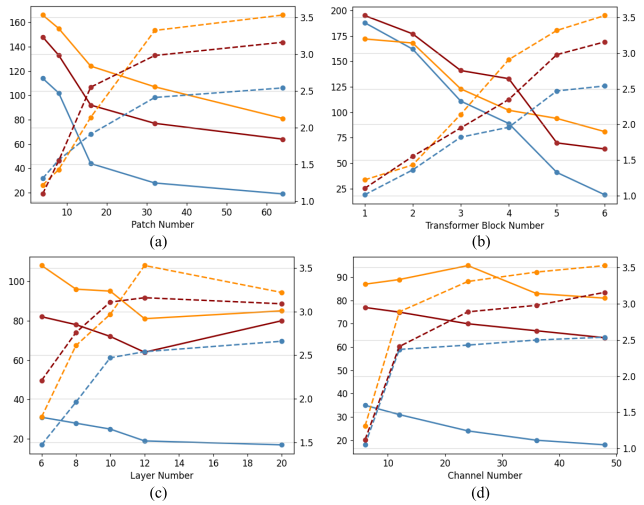

Fig. 10. Ablation study on different settings of SwinIT architecture. Quantitative results with FID (left vertical axis) and IS (right) metrics are tested on style transfer tasks. Different datasets are distinguished by color, and day2night (photo2art, summer2winter) is marked with blue (orange, red).

the consideration of fine-grained patch dependencies helps increase the model's performance.

b) Selection of transformer block number, layer number, and channel number: We show the effects of the number of Transformer blocks (B) in the generator, the total number of layers (L), and the number of channels (C) for the first Transformer block on model performance in Figs. 10(b),(c),(d), respectively. It is observed that the FID and IS are positively correlated with these three hyperparameters, basically. It can be noted that the performance keeps increasing for more hierarchical blocks but is limited by the number of downsamplings allowed at the determined resolution of the image. When it comes to the layer and channel numbers, performance gains tend to be saturated gradually, along with a fairly high increase in the number of parameters. Therefore, to balance the performance and model size, we choose $B = 6, L = 12, C = 48$ in the rest of the experiments. We choose $B = 6$ for both the content and style encoders to obtain a relatively small model.

I. Analysis

1) Variations: To verify the role of cycle consistency, we conduct experiments to evaluate the performance of our method with or without cycle consistency loss. Model G in the ablation study is taken as the baseline. We perform the semantic image synthesis and multimodal translation tasks in inverse settings (i.e., adding image → mask or night → day mapping processes). Without modifying the architecture of SwinIT, we tune the cycle consistency loss weights and make the model converge as much as possible. Visual validation results are shown in Fig. 9. The proposed method with cycle consistency shows worse results than the baseline. Mask → image mapping on Cityscapes produces blurry results and loses too many details of the ground truth image. Day → night mapping introduces irregular artifacts, and there are areas where the style is not fully transferred.

2) Limitations: In our experiments, we assume the translation tasks are injections, which causes the model to fail to perform inverse mapping well. In the pursuit of high-quality image generation, we need to trade model generality for expense. As image translation tasks increase, the number of generators and discriminators will increase accordingly. Therefore, in addition to model compression, the I2I translation task for universal domains is an open problem. Specifically, to accommodate different domain mappings in a single model, we can use paired head-tail architectures for each mapping task separately. The content and style images are encoded into the latent space by head, and then the I2I translation in the latent space is done using the proposed framework, and finally the translated latent encoding is mapped back to the image through the tail encoding. In addition, task encodings can be embedded in the generator to prevent mutual interference between domain mappings.

V. Conclusion

In this paper, we introduce SwinIT, a hierarchical Transformer-based framework for various I2I translation tasks without cycle consistency constraints. We explore adaptive denormalization and normalization strategies to effectively capture and fuse content and style representations in a coarse-to-fine manner. Moreover, we design a wavelet transformation matching loss to recover rich details. Extensive experiments demonstrate the superiority of our proposed method in several diverse I2I translation tasks. Hopefully, this paper will be able to inspire researchers to design a unified framework for versatile tasks.

References

[1] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2332–2341.

[2] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8185–8194.

[3] C. Zhang et al., "Density-aware haze image synthesis by self-supervised content-style disentanglement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4552–4572, Jul. 2022.

[4] Y. Gao, S. Ma, and J. Liu, "DCDR-GAN: A densely connected disentangled representation generative adversarial network for infrared and visible image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 549–561, Feb. 2023.

[5] H. Wu et al., "Multi-grained attention networks for single image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 512–522, Feb. 2021.

[6] J. Zhang et al., "A two-stage attentive network for single image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1020–1033, Mar. 2022.

[7] Z. Liu, Z. Li, X. Wu, Z. Liu, and W. Chen, "DSRGAN: Detail prior-assisted perceptual single image super-resolution via generative adversarial networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7418–7431, Nov. 2022.

[8] Y. Gao et al., "Wallpaper texture generation and style transfer based on multi-label semantics," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1552–1563, Mar. 2022.

[9] Y. Deng et al., "StyTr2: Image style transfer with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11316–11326.

[10] X. Zhong, T. Lu, W. Huang, M. Ye, X. Jia, and C.-W. Lin, "Grayscale enhancement colorization network for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1418–1430, Mar. 2022.

[11] Y. Zhao et al., "VCGAN: Video colorization with hybrid generative adversarial network," 2021, *arXiv:2104.12357*.

[12] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2868–2876.

[13] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[14] D. S. Tan, Y.-X. Lin, and K.-L. Hua, "Incremental learning of multi-domain image-to-image translations," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1526–1539, Apr. 2021.

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

[16] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 172–189.

[17] H.-Y. Lee et al., "DRIT++: Diverse image-to-image translation via disentangled representations," *Int. J. Comput. Vis.*, vol. 128, nos. 10–11, pp. 2402–2417, Nov. 2020.

[18] Y. Gao et al., "High-fidelity and arbitrary face editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16110–16119.

[19] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 319–345.

[20] D. Ruta et al., "ALADIN: All layer adaptive instance normalization for fine-grained style similarity," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11906–11915.

[21] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[22] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[23] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[24] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7242–7252.

[25] S. Kim, J. Baek, J. Park, G. Kim, and S. Kim, "InstaFormer: Instance-aware image-to-image translation with transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18300–18310.

[26] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1510–1519.

[27] L. Jiang, C. Zhang, M. Huang, C. Liu, J. Shi, and C. C. Loy, "TSIT: A simple and versatile framework for image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020.

[28] H. Fu, T. Yu, X. Wang, and H. Ma, "Cross-granularity learning for multi-domain image-to-image translation," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3099–3107.

[29] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1–9.

[30] Y. Zhang, S. Liu, C. Dong, X. Zhang, and Y. Yuan, "Multiple cycle-in-cycle generative adversarial networks for unsupervised image super-resolution," *IEEE Trans. Image Process.*, vol. 29, pp. 1101–1112, 2020.

[31] W. Zhang, Y. Liu, C. Dong, and Y. Qiao, "RankSRGAN: Generative adversarial networks with ranker for image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3096–3105.

[32] J. Kim, M. Kim, H. Kang, and K. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," 2019, *arXiv:1907.10830*.

[33] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[34] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[35] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[36] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[37] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5790–5799.

[38] R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, and X. Luo, "MADNet: A fast and lightweight network for single-image super resolution," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1443–1453, Mar. 2021.

[39] J. Liu, W. Zhang, Y. Tang, J. Tang, and G. Wu, "Residual feature aggregation network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2356–2365.

[40] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, "Deep burst super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9205–9214.

[41] R. Zhang et al., "Real-time user-guided image colorization with learned deep priors," 2017, *arXiv:1705.02999*.

[42] S. Wan, Y. Xia, L. Qi, Y.-H. Yang, and M. Atiquzzaman, "Automated colorization of a grayscale image with seed points propagation," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1756–1768, Jul. 2020.

[43] Z. Xu, T. Wang, F. Fang, Y. Sheng, and G. Zhang, "Stylization-based architecture for fast deep exemplar colorization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9360–9369.

[44] H. Fu, C. Tian, X. Wang, and H. Ma, "Stacked semantically-guided learning for image de-distortion," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4519–4527.

[45] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, "Contextual residual aggregation for ultra high-resolution image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7505–7514.

[46] N. Wang, Y. Zhang, and L. Zhang, "Dynamic selection network for image inpainting," *IEEE Trans. Image Process.*, vol. 30, pp. 1784–1798, 2021.

[47] M. Tomei, M. Cornia, L. Baraldi, and R. Cucchiara, "Art2Real: Unfolding the reality of artworks via semantically-aware image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5842–5852.

[48] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.

[49] S. Liu et al., "AdaAttN: Revisit attention mechanism in arbitrary neural style transfer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6629–6638.

[50] J.-Y. Zhu et al., "Toward multimodal image-to-image translation," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 465–476.

[51] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[52] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1857–1865.

[53] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 694–711.

[54] X. Yu, Y. Chen, S. Liu, T. Li, and G. Li, "Multi-mapping image-to-image translation via learning disentanglement," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–11.

[55] Z. Shen, M. Huang, J. Shi, X. Xue, and T. S. Huang, "Towards instance-level image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3678–3687.

[56] J. Lin, Z. Chen, Y. Xia, S. Liu, T. Qin, and J. Luo, "Exploring explicit domain supervision for latent space disentanglement in unpaired image-to-image translation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1254–1266, Apr. 2021.

[57] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[58] Y. Wang, Z. Zhang, W. Hao, and C. Song, "Multi-domain image-to-image translation via a unified circular framework," *IEEE Trans. Image Process.*, vol. 30, pp. 670–684, 2021.

[59] N. Parmar et al., "Image transformer," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4055–4064.

[60] H. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12294–12305.
[61] Y. Jiang, S. Chang, and Z. Wang, "Transgan: Two pure transformers can make one strong gan, and that can scale up," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021.
[62] B. Zhang et al., "StyleSwin: Transformer-based GAN for high-resolution image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11294–11304.
[63] S. Mo, M. Cho, and J. Shin, "InstaGAN: Instance-aware image-to-image translation," 2018, *arXiv:1812.10889*.
[64] Z. Liu et al., "Video Swin transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3192–3201.
[65] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.
[66] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[67] Q. Mao, H.-Y. Tseng, H.-Y. Lee, J.-B. Huang, S. Ma, and M.-H. Yang, "Continuous and diverse image-to-image translation via signed attribute vectors," *Int. J. Comput. Vis.*, vol. 130, no. 2, pp. 517–549, Feb. 2022.
[68] X. Hu, X. Zhou, Q. Huang, Z. Shi, L. Sun, and Q. Li, "QS-attn: Query-selected attention for contrastive learning in I2I translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18270–18279.
[69] J. Choi, D. Kim, and B. C. Song, "Style-guided and disentangled representation for robust image-to-image translation," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 1, pp. 463–471.
[70] Y. Zhang et al., "Domain enhanced arbitrary image style transfer via contrastive learning," in *Proc. ACM SIGGRAPH Conf.*, 2022, pp. 1–8.
[71] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1520–1529.
[72] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[73] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*.
[74] F. Yu et al., "BDD100K: A diverse driving dataset for heterogeneous multitask learning," 2018, *arXiv:1805.04687*.
[75] R. Tylecék and R. Sára, "Spatial pattern templates for recognition of objects with regular structure," in *Proc. GCPR*, Saarbrucken, Germany, 2013, pp. 364–374.
[76] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[77] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5122–5130.
[78] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–12.
[79] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[80] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 2234–2242.
[81] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[82] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1–9.
[83] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[84] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 418–434.

Jin Liu received the B.S. degree from North China Electric Power University, China, in 2020. He is currently pursuing the Ph.D. degree with the Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include deep learning and machine learning and their applications to computer vision, especially image and video synthesis and low-quality enhancement.

Huiyuan Fu (Member, IEEE) received the Ph.D. degree in computer science from the Beijing University of Posts and Telecommunications in 2014. He is a Professor with the School of Computer Science, Beijing University of Posts and Telecommunications. He has published more than 60 papers in journals (such as IEEE TRANSACTIONS) or conferences (such as IEEE CVPR, IEEE ICCV, ACM Multimedia, and AAAI). His research area includes visual big data, machine learning, pattern recognition, and multimedia systems. He received the Best Student Paper Award at IEEE ICME in 2016 and the Best Paper Award Runner-Up at IEEE/ACM ICDSC in 2014.

Xin Wang (Senior Member, IEEE) received the B.S. degree in telecommunications engineering and the M.S. degree in wireless communications engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 1990 and 1993, respectively, and the Ph.D. degree in electrical and computer engineering from Columbia University, New York, NY, USA. She was a member of Technical Staff, Bell Labs Research, Lucent Technologies, NJ, USA, in the area of mobile and wireless networking; and an Assistant Professor with the Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY. She is currently an Associate Professor with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY. Her research interests include wireless networks, mobile computing, big data, and machine learning. She was a recipient of the NSF Career Award in 2005 and the ONR Challenge Award in 2010. She has served in executive committee and technical committee for numerous conferences and funding review panels. She was an Associate Editor of IEEE TRANSACTIONS ON MOBILE COMPUTING.

Huadong Ma (Fellow, IEEE) received the B.S. degree in mathematics from Henan Normal University, Xinxiang, China, in 1984, the M.S. degree in computer science from the Shenyang Institute of Computing Technology, Chinese Academy of Science, Beijing, China, in 1990, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Science, in 1995. He is currently a Professor with the School of Computer Science, Beijing University of Posts and Telecommunications, China. From 1999 to 2000, he held a visiting position with the University of Michigan, Ann Arbor, MI, USA. He has published more than 300 papers in journals (such as ACM/IEEE TRANSACTIONS) and conferences (such as IEEE CVPR, IEEE ICCV, ACM MobiCom, and ACM SIGCOMM) and five books. His current research interests include the Internet of Things, sensor networks, and multimedia computing. He received the Natural Science Award of the Ministry of Education, China, in 2017. He also received the 2019 Prize Paper Award of IEEE TRANSACTIONS ON MULTIMEDIA, the Best Paper Award from IEEE MULTIMEDIA, the Best Paper Award in IEEE ICPADS 2010, and the Best Student Paper Award in IEEE ICME 2016 for his coauthored papers. He received the National Funds for Distinguished Young Scientists in 2009. He serves as the Chair for ACM SIGMOBILE China. He was/is an Editorial Board Member of IEEE TRANSACTIONS ON MULTIMEDIA, IEEE INTERNET OF THINGS JOURNAL, and *ACM Transactions on Internet of Things*.