Hierarchy Flow For High-Fidelity Image-to-Image Translation

Weichen Fan[†], Jinghuan Chen[†], Ziwei Liu, *Member, IEEE,*

Abstract—Image-to-image (I2I) translation comprises a wide spectrum of tasks. Here we divide this problem into three levels: strong-fidelity translation, normal-fidelity translation, and weak-fidelity translation, indicating the extent to which the content of the original image is preserved. Although existing methods achieve good performance in weak-fidelity translation, they fail to fully preserve the content in both strong- and normal-fidelity tasks, e.g. sim2real, style transfer and low-level vision. In this work, we propose *Hierarchy Flow*, a novel flow-based model to achieve better content preservation during translation. Specifically, *1)* we first unveil the drawbacks of standard flow-based models when applied to I2I translation. *2)* Next, we propose a new design, namely hierarchical coupling for reversible feature transformation and multi-scale modeling, to constitute Hierarchy Flow. *3)* Finally, we present a dedicated aligned-style loss for a better trade-off between content preservation and stylization during translation. Extensive experiments on a wide range of I2I translation benchmarks demonstrate that our approach achieves state-of-the-art performance, with convincing advantages in both strong- and normal-fidelity tasks. Code and models will be at https://github.com/WeichenFan/HierarchyFlow.

Index Terms—Image-to-image translation, generative model, normalizing flow, low-level vision.

1 Introduction

Image-to-image translation [9] is a long-standing topic in computer vision, which is required to learn a mapping between two different visual domains while preserving the semantic information (content) of the source domain and obtaining the domain properties (style) of the target domain. Many applications, such as neural style transfer [10]–[13], super-resolution [14], [15], image enhancement [16] and photo-realistic synthesis [17]–[19], can be formulated as I2I translation problems. Among most tasks, fully preserving semantic information during translation is important yet challenging, especially in scenarios where the content gap between source and target domains is large. According to the requirement of content preservation during translation, we further divide these tasks into three levels: strong-fidelity translation, normal-fidelity translation, and weak-fidelity translation (see Figure 2). In this work, we are interested in strong- and normal-fidelity settings, where content preservation plays a crucial role during translation.

Existing I2I translation methods can be broadly categorized into two approaches. Some methods learn a bijective mapping between source and target images, by forcing the translated images to be reconstructed back to the source images during training using a cyclic loss (e.g. [1]). Others try to fully disentangle content and style information from an image and achieve image translation by switching style information between source and target images (e.g. [20]). However, both approaches suffer from different levels of content distortion in translated images, since cyclic loss and feature disentanglement usually failed when rich and complex semantic information is required to be preserved, especially in strong- and normal-fidelity translation tasks. Relatively few methods focus on addressing this problem directly, and most among them suffer from carefully-designed tricks and are unable to generalize well to a wide range of tasks, since they either require auxiliary inputs such as paired images or additional information guidance, etc. (e.g. [19]), or utilize complex contrastive training or pre-trained tasks aiming for better feature disentanglement learning (e.g. [21], [22]).

Content preservation remains a challenging problem in I2I translation. We consider flow-based models, also called normalizing flow, a subclass of deep generative networks that learns the exact likelihood of data distribution through a chain of basic blocks with fully-reversible transformations, which can be a perfect fit in the requirement of content preservation in image generation. ArtFlow [23] is the first work to use the flow-based model in I2I translation, specifically in style transfer task only. It proves the superiority of flow-based models in addressing the "content leakage" problem through lossless and unbiased feature extraction and image reconstruction. However, although ArtFlow achieves better content preservation compared to other methods, it suffers from severe checkerboard artifacts problem in the translated images (see Figure 4). We further investigate the checkerboard issue and finally identify its root cause as the squeeze operation that is widely-used in flow-based models for multi-scale architecture [24], [25]. Sec.3.1 shows more analyses in detail. Therefore, we focus on designing a new framework that can utilize the superiority of flow-based models in content preservation for I2I translation, and also avoid the checkerboard artifacts problem as in ArtFlow.

In this work, we propose *Hierarchy Flow*, which is a new flow-based model dedicated to unpaired I2I translation with good content-preserving ability. To avoid the problematic squeeze operation in multi-scale architecture for flow-based

- *Weichen Fan, Jinghuan Chen and Ziwei Liu are with S-Lab, Nanyang Technological University.*
E-mail: fanweichen2383@gmail.com, chenjh1997@yahoo.com, ziwei.liu@ntu.edu.sg
- *† Equal contribution.*

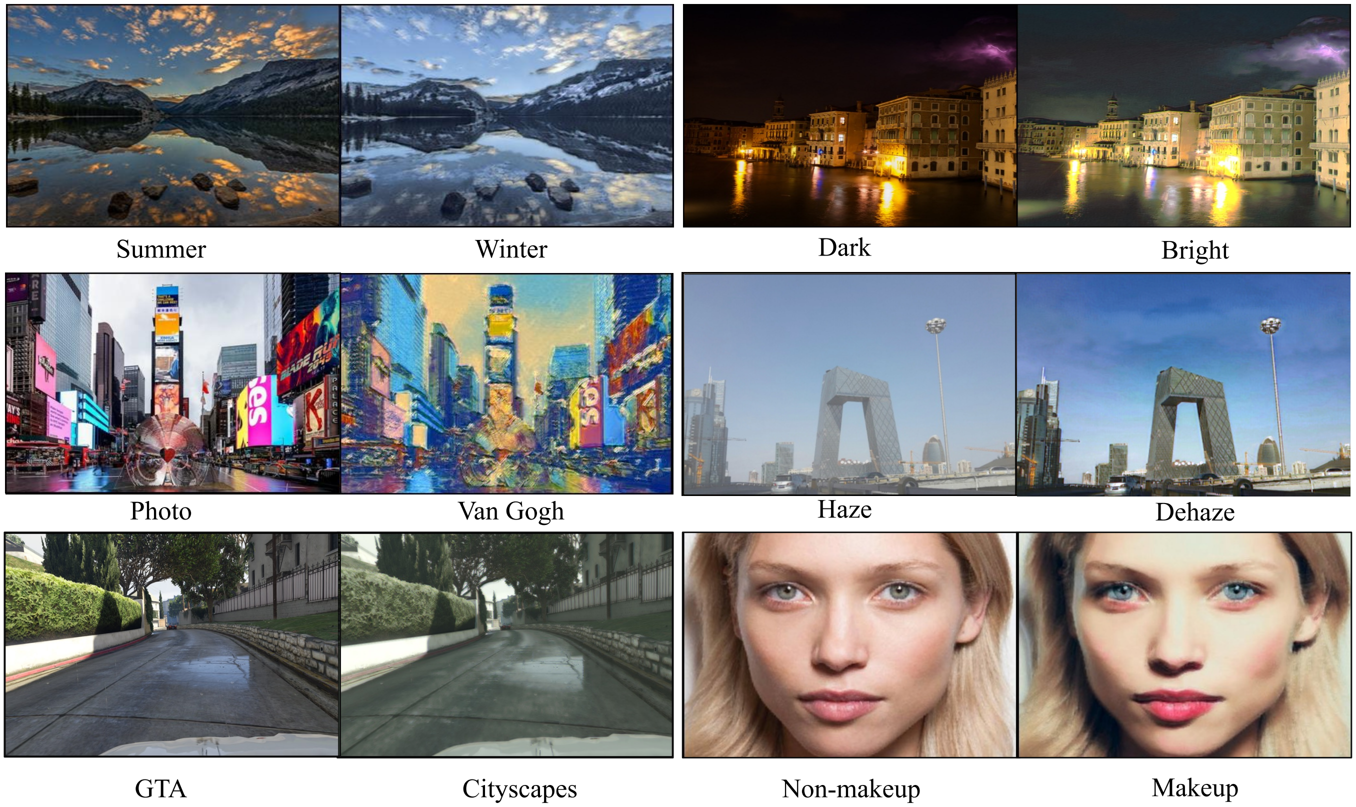

Fig. 1. Given two images in different visual domains, our model learns to translate from one to the other with high fidelity in various tasks. (From top-left to bottom-right: 1. Summer to Winter [1]; 2. Dark to Bright [2]; 3. Photo to Van Gogh [3]; 4. Hazy to Clear [4], [5]; 5. GTA [6] to Cityscapes [7]; 6. Non-makeup to makeup [8].

models, we present a novel basic block design, named *Hierarchical Coupling Layer*, for efficient feature transformation and multi-scale modelling. In our model, feature extraction is done in a hierarchical way which can gradually remove style-specific features by a series of subtractive coupling operations in the forward pass. After feature extraction, we use Adaptive Instance Normalization [12] to perform transformation upon deep features by replacing the statistical information (mean/std vector of features) of source features with those of target features. Finally, translated images are generated through the reversed pass of the network. Following most I2I translation tasks especially in style transfer [12], content loss and style loss calculated based on a pre-trained VGG encoder [26] are adopted. We further extend the idea of style loss and introduce its simple extension named *aligned-style loss*, which takes the trade-off between content preservation and stylization into consideration, to further improve translation results, especially in scenarios where content gap is large between source and target domains, e.g. GTA [6] to Cityscapes [7] translation.

We apply the proposed framework to a wide range of applications, and plausible results (see Figure 1) indicate the significance and effectiveness of the design in our method. To the best of our knowledge, we are the first I2I translation work that evaluates on both high-level (e.g. GTA to Cityscapes) and low-level (e.g. Low-light enhancement) vision tasks and achieves superior results in both areas. We summarize the contributions of this work as below: **1)** We divide image-to-image translation tasks into three subsets: strong-, normal- and weak-fidelity translation, according to the requirement of content preservation. **2)** We unveil the main drawback of flow-based models in I2I translation tasks, and propose Hierarchy Flow, a novel design for unpaired high-fidelity image-to-image translation. **3)** We design a novel aligned-style loss for efficient content-preserving feature transformation. **4)** We demonstrate that Hierarchy Flow outperforms previous methods with high

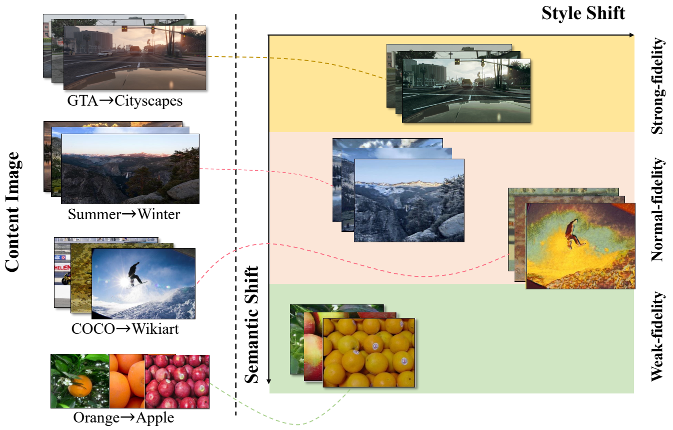

Fig. 2. Illustration of three levels of image-to-image translation tasks: strong-, normal- and weak-fidelity translation, where the requirement of content preservation decreases gradually.

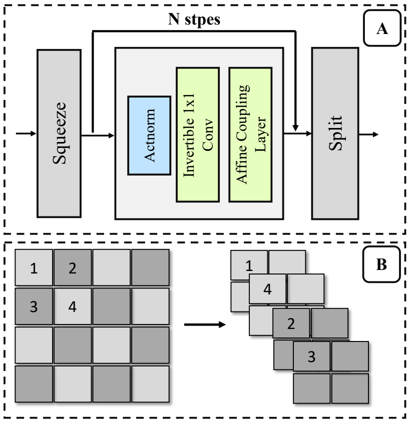

Fig. 3. (A). The basic components of GLOW [25] consist of a squeeze operation followed by a series of invertible layers for non-linear transformation. (B). The squeeze operation [24] reorganizes the features map following a spatial checkerboard pattern.

fidelity and vivid stylization in extensive experiments.

2 Related Work

2.1 Image-to-Image Translation

Generic Image-to-Image Translation: The previous generic I2I models [20], [27], [28] suffer from the problem of content distortion, even though a lot of regularization methods have been proposed to reduce the impact, including cyclic consistency, contrastive learning, etc. For weak-fidelity translations, where the content can be heavily modified, these methods are appropriate, while for strong- and normal-fidelity translations, they do not perform well. Our proposed model can be applied to both normal- and strong-fidelity translations without the problem of content distortion.

Strong-fidelity Translation: Strong-fidelity I2I translation means the content of the source image should be preserved to a great extent. Sim2real translation [19], [22], colorization [29], and low-level visions such as low light enhancement [2], [16], raindrop removal [30], [31] and dehazing [31]–[33] belong to the strong-fidelity setting. These problems require the translated images to retain the exact rich and complex semantic information in source images. To achieve high content preservation, previous methods require paired training images or auxiliary inputs such as semantic segmentation masks. Most recently, VSAIT [34] proposes a new framework based on vector symbolic architectures to directly solve "semantic flipping" problem and achieve current SOTA results in unpaired I2I translation.

Normal-fidelity Translation: Normal-fidelity I2I translation includes style transfer [10], [12], [13], [23], [35], [36], season and weather transfer [37], etc. In this setting, the source and target domains usually show different visual effects, such as weather conditions and artistic styles, but share similar structural information, the primary objective is to transfer the overall visual effects of source domains to match those in target domains. Previous work have shown plausible overall visual results in these tasks, while a certain level of content distortion can be found when we zoom in to the details of translated images.

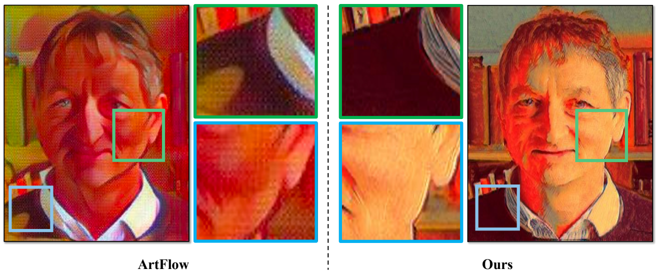

Fig. 4. The checkerboard artifacts. Left: output sample from ArtFlow [23]; Right: our result. Compared to the output of ArtFlow which suffers from checkerboard artifacts (shown in zoomed-in boxes), we generate smooth images with high fidelity.

Weak-fidelity Translation: Weak-fidelity I2I translation refers to problems where the source and target images may lie in completely different domains or modals, the translation is to be performed on a high semantic level, which means the content information of source images can be modified a lot. Label to image [38] and object to object translation [1], [28], [39], [40] belong to this type of problem.

2.2 Normalizing Flow

Normalizing flow is a type of generative model that uses a sequence of invertible mappings to transform from distribution to distribution, and it is accurate and efficient in both density estimation and sampling [41]. Dinh et al. [42] first propose a flow-based generative model, NICE. After that, GLOW [25], RealNVP [24], and FLOW++ [43] are proposed to improve the sample efficiency and density estimation performance. More recently, BeautyGLOW [44] is proposed for makeup transfer. Besides, ArtFlow [23] proves that the normalizing flow is unbiased in neural style transfer compared with the previous work.

3 Our Approach

In this section, we first give a brief introduction of flow-based generative models and unveil its main drawback in I2I translation, which is the checkerboard artifacts, in Sec.3.1; next, we introduce the design of our proposed Hierarchy Flow in details in Sec.3.2, which solves the checkerboard issues of previous methods and achieve better content preservation in high-fidelity image translation.

3.1 Preliminary

3.1.1 *Flow-based Generative Model*

Flow-based model is a subset of generative models that learns the exact log-likelihood of a high dimensional data distribution through a sequence of fully reversible transformations. Let x be a high-dimensional variable with unknown distribution $x \sim p(x)$, a generative model $p_\theta(x)$ with parameters θ is designed to estimate distribution $p(x)$ given dataset $\mathcal{D}$, with training objective $min\ \mathcal{L}(\mathcal{D})$, where

$$\mathcal{L}(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N} -log(p_\theta(x^{(i)})) \tag{1}$$

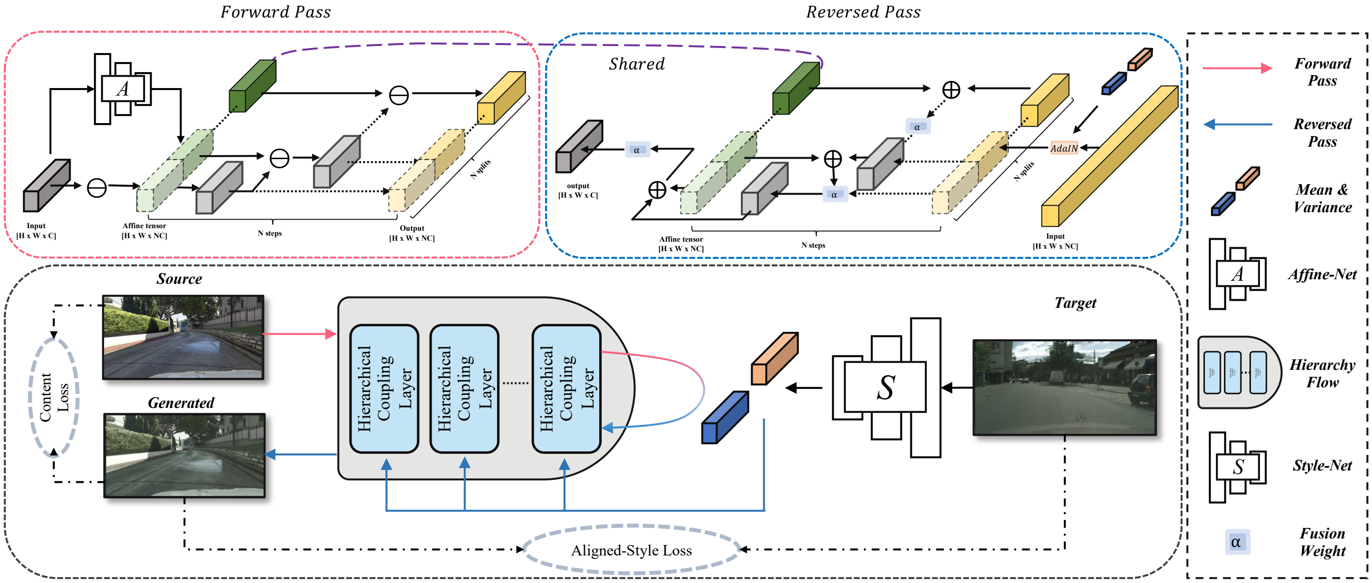

Fig. 5. The framework of our proposed Hierarchy Flow. Given a source image and a target image, the source image is encoded by *Hierarchy Flow* in the forward pass represented by the red arrows, then fused with style information extracted from the target image by *Style-Net* in the form of AdaIN [12], and finally reverse-decoded by *Hierarchy Flow* and denoted in blue arrows to generate translated image.

Most flow-based generative models [24], [25], [42] formulate the generative process as

$$x \overset{f_1}{\longleftrightarrow} h_1 \overset{f_2}{\longleftrightarrow} h_2 \cdots \overset{f_k}{\longleftrightarrow} z \qquad (2)$$

where z is latent variable, $f_\theta = f_1 \circ f_2 \circ \cdots \circ f_k$ is a sequence of invertible functions such that z and x satisfy the relationship $z = f_\theta(x)$ and $x = f_\theta^{-1}(z)$. One of the most frequently-used flow-based networks for image synthesis is GLOW [25]. As shown in Figure 3, it combines a series of steps of flow in a multi-scale architecture with squeeze operations [24] between each scale to effectively transform feature scale by trading spatial size for channel size. In each step of flow, features are transformed by actnorm activation followed by an invertible 1x1 convolution and lastly an affine coupling layer. Previous work ArtFlow [23] follows the network design of GLOW to achieve reversible transformations in universal style transfer, and solve the content leakage problem by lossless and unbiased image projection and reversion.

3.1.2 Checkerboard Artifacts Problem

The squeeze operation plays a vital role in GLOW for multi-scale feature transformation. As shown in Figure 3(B), it follows a spatial checkerboard pattern to transform a $H \times W \times C$ tensor (left) into $\frac{H}{2} \times \frac{W}{2} \times 4C$ tensor (right), to efficiently implement multi-scale architecture by trading spatial size for channel size. During reversed pass of the model, one can easily “undo” the squeeze operation by restoring the spatial dimension of the tensors.

In ArtFlow, the style feature transfer module (AdaIN [12] or WCT [13]) is applied to the encoded features before the reversed pass to generate translated images. As a result, significant changes have been made to each individual channel of the features, and spatial misalignment will be produced when unsqueeze operations are performed on the transferred features during the reversed pass, leading to the obvious checkerboard artifacts in the output samples (see Figure 4: Left).

In order to utilize the usage of invertible network design of flow-based models and to solve checkerboard artifacts problem in ArtFlow, we aim to re-design the squeeze operation for multi-scale architecture to achieve content-fixed and artifacts-free image-to-image translation.

3.2 Hierarchy Flow

As shown in Figure 5, we propose Hierarchy Flow, which is a flow-based model with a novel design of basic block named *Hierarchical Coupling Layer*. In general, given a set of images (I_s, I_t), a series of hierarchical coupling layers encode the source image I_s to obtain the source features in the forward network inference. The target image I_t is fed into a *Style-Net* to obtain the style features. After that, we use AdaIN [12] to perform style feature transfer to fuse the source features and style features, and finally perform image reconstruction through the reversed pass of the network to generate a translated image. In our model, the network architecture is carefully designed to be fully reversible. Therefore, combined with AdaIN, we can achieve I2I translation with desired content preservation.

3.2.1 Hierarchical Coupling Layer

By combining squeeze operation with affine coupling layer together, *hierarchical coupling layer* enables complex feature transformation and multi-scale modeling inside one single block without spatial squeezing. Instead, we use hierarchical subtraction along the channel dimensions to implement spatial feature fusion and transformation in a learnable manner. Algorithms 1 and 2 show the details of the forward and reversed pass respectively.

A. Forward Pass. Given an input tensor x with dimension

Algorithm 1 Forward Pass.

```
FORWARD($x$)
    $a = \textbf{Affine-Net}(x)$
    $a_1, a_2, \cdots, a_n = \textbf{split}(a)$
    $h_1 = x - a_1$
    $h_i = h_{\mathbf{i}-1} - a_{\mathbf{i}}$  for  $\mathbf{i} \leftarrow 2, n$
    $y = \textbf{concat}(h_1, h_2, \cdots, h_n)$
    return y
```

Algorithm 2 Reversed Pass.

```
REVERSED($y$, $a_{1\cdots n}$, style feature $\mu$, $\sigma$)
    $y = \textbf{AdaIN}(y, \mu, \sigma)$
    $y_1, y_2, \cdots, y_n = \textbf{split}(y)$
    $h_n = y_n + a_n$
    $h_{\mathbf{i}} = \alpha \cdot (y_{\mathbf{i}} + a_{\mathbf{i}}) + (1 - \alpha) \cdot h_{\mathbf{i}+1}$  for  $\mathbf{i} \leftarrow n - 1, 1$
    $x = h_1$
    return x
```

$[H \times W \times C]$, we first apply an affine transformation with an *Affine-Net* which can be any neural network, where the input tensor undergoes a channel-wise expansion and is mapped to an affined tensor with dimension $[H \times W \times nC]$ with expansion rate n. With the affine tensor separated into n splits along channel dimension, we then apply a hierarchical subtractive coupling for x in n-steps, and obtain output y by concatenating the n intermediate feature maps.

B. Reversed Pass. To compute the inverse of the above transformation, we can simply apply n steps of addictive coupling between input tensor y and affine tensor a, and fuse the n intermediate feature maps to obtain output tensor x. To better facilitate the fusion process in the training, we apply a learnable fusion weight α in each step of fusion, which measures the importance of each split of features during spatial fusion and transformation adaptively.

With the design of hierarchical coupling, we enable multi-scale feature transformation and fusion inside each basic block. Therefore, we can easily stack multiple blocks directly to implement more complex network modeling, without spatial squeezing as in previous flow-based methods, since adaptive spatial fusion has been applied inside each block. Despite of its simplicity in network design, Hierarchy Flow shows great improvement in high-fidelity translation tasks with better content preservation and artifacts-free image generation, which indicates the effectiveness and significance of our proposed method.

3.2.2 Style-Net and AdaIN

Our *Style-Net* follows the design of Style-Encoder in MUNIT [39], which consists of a series of convolutional layers with stride 2 followed by a global average pooling and two linear layers that output a mean vector μ and a variance vector σ. The purpose of the style network is to extract style information, thus we do not use any normalization layers in the network, which would modify the style information.

AdaIN is first proposed in [12], it separates deep features into normalized feature map and mean/std vectors, which can be referred as content and style information respectively. To perform style feature transfer, it first scales normalized source feature x by variance vector σ, then shift it with mean vector of μ, where σ and μ are the outputs of *Style-Net*:

$$\mathrm{AdaIN}(x, \mu, \sigma) = \sigma(\frac{x - \mu(x)}{\sigma(x)}) + \mu \tag{3}$$

In our model, AdaIN is applied to every hierarchical coupling layer before the reversed pass.

3.2.3 Loss Function

Our objective function can be expressed as:

$$\mathcal{L} = \mathcal{L}_c + \lambda \mathcal{L}_{as} \tag{4}$$

where L_c is the content loss, L_{as} is our proposed aligned-style loss, and λ is a weighting factor used to trade-off between content and style.

Content Loss: Following [12], the content loss is defined as the Euclidean distance between the channel-wise normalization of VGG features for the generated image $\hat{x}$ and the source image x.

$$\mathcal{L}_c = \|norm(\phi(\hat{x})) - norm(\phi(x))\|_2 \tag{5}$$

where ϕ refers to the layer $relu\ 4\_1$ of a pre-trained VGG-19 encoder, $norm$ denotes the channel-wise normalization.

Aligned-Style Loss: Considering that the semantic information extracted from VGG-19 of the source image and the target image are not exactly matched in unpaired translation, we extend the style loss in [12] to be *aligned-style loss* by setting a parameter k to adjust the percentage of extracted tensors that are used for loss computation. We define S as an ascending sort function. Given a source image x, a target image y, and the transferred image $\hat{x}$, with an energy function $E(\phi_i(\hat{x}), \phi_i(y)) = \|\mu(\phi_i(\hat{x})) - \mu(\phi_i(y))\|_2$, where ϕ_i $(i \in L = \{1, 2, 3\})$ represents a set of pre-trained VGG-19 layers $\{relu1\_1, relu2\_1, relu3\_1\}$, we could have the chosen index:

$$\mathcal{C} = \{c \in \mathbb{N}_S | c \leqslant kN, 0 < k \leqslant 1\} \tag{6}$$

where $\mathbb{N}_S$ is the set of indexes of the sorted tensor $S(E(\phi_i(\hat{x}), \phi_i(y)))$, N denotes its total length, and k is the weighting parameter. Therefore,

$$\begin{aligned}\mathcal{L}_{as} = \sum_{i=1}^{L} \sum_{j \in C} \|\mu(\phi_i(\hat{x})_j) - \mu(\phi_i(y)_j)\|_2 + \\ \sum_{i=1}^{L} \sum_{j \in C} \|\sigma(\phi_i(\hat{x})_j) - \sigma(\phi_i(y)_j)\|_2\end{aligned} \tag{7}$$

where $\phi_i(x)_j$ denotes the j^{th} channel of the output tensor of the i^{th} layer from the set $\{relu1\_1, relu2\_1, relu3\_1\}$ of a pre-trained VGG-19 encoder.

4 Experiments

To demonstrate the effectiveness of our method in normal- and strong-fidelity translation tasks, we show comparisons between our proposed Hierarchy Flow and other state-of-the-art methods of respective fields in this section. More results can be found in supplementary materials.

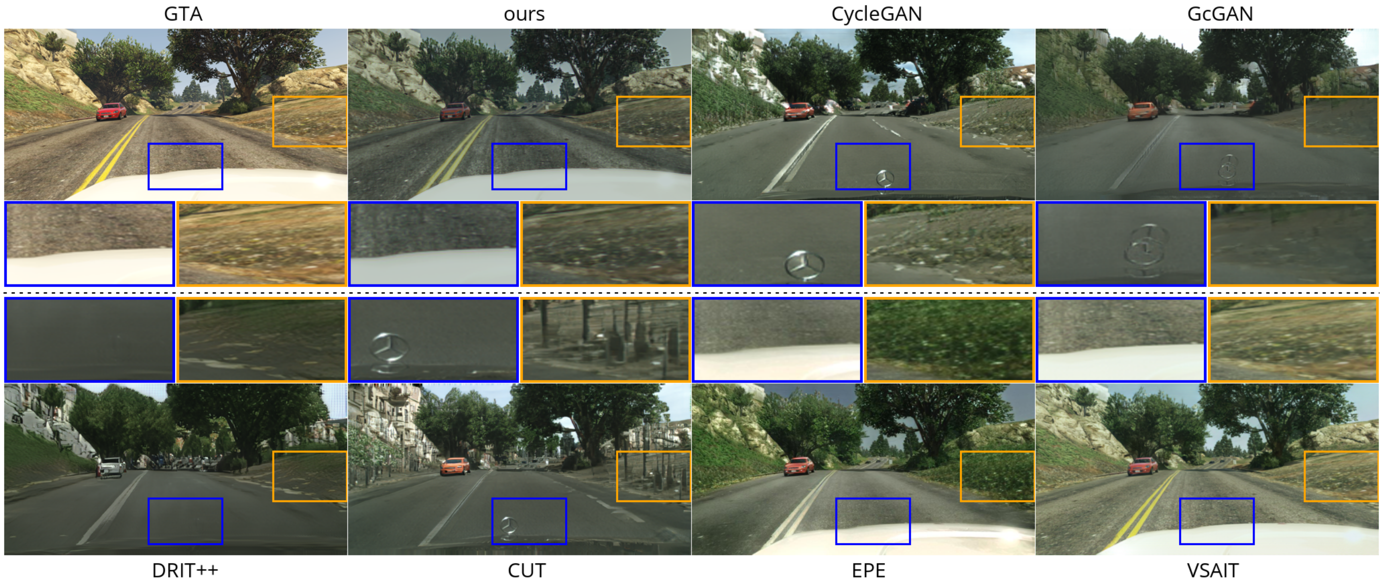

Fig. 6. Visual results of GTA to Cityscapes compared with the state-of-the-art methods. CycleGAN [1], GcGAN [45] and CUT [40] hallucinate logos (blue box), while DRIT++ [46] and EPE [19] generate extra grass on the ground (orange box). Compared to VSAIT [34], our model generates better Cityscapes-style images.

TABLE 1
Quantitative evaluation on GTA to Cityscapes. Metrics include average pixel prediction accuracy (pxAcc), average class prediction accuracy (clsAcc), mean IoU (mIoU), and SSIM and FID. We achieve best results with smallest model size.

Method	pxAcc↑	clsAcc↑	mIoU↑	SSIM↑	FID↓	Params↓	FLOPs↓
CycleGAN [1]	75.93	39.12	28.92	0.70	22.12	22.76M	454.91G
GcGAN [45]	70.26	38.13	27.34	0.67	12.32	7.84M	84.71G
DRIT++ [46]	75.89	35.44	27.31	0.55	14.69	17.08M	170.18G
CUT [40]	70.81	37.12	26.43	0.61	21.18	11.38M	128.26G
ArtFlow [23]	76.05	45.37	32.15	0.52	8.59	6.42M	90.2G
VSAIT [34]	75.33	42.23	30.33	**0.93**	7.94	11.38M	128.26G
ours	**79.63**	**45.67**	**33.76**	0.87	**7.87**	**0.68M**	**10.13G**

4.1 Experimental Setup

Network Architecture. As mentioned in Sec 3.2, a stacked sequence of basic blocks can be used for complex modeling in different tasks. In the following experiments, we introduce 4 variants of model size with different number of basic blocks for different tasks, which includes:

1) *HF*: 2 blocks with expansion rates $[10, 4]$ (Sec. 4.2);
2) *HF+*: 3 blocks with expansion rates $[4, 5, 2]$ (Sec. 4.3);
3) *HF++*: 3 blocks with expansion rates $[10, 4, 4]$;
4) *HF*†: 4 blocks with expansion rates $[10, 4, 4, 4]$ (Sec. 4.4).

More studies of the network structure can be found in the ablation study.

Affine-Net Design. Inside each basic block, *Affine-Net* is defined as a 3-layer perceptron with *Conv-IN-ReLU-Conv-IN-ReLU-Conv-ReLU* in specific, where the first two convolutional layers double the input channel dimension and the last one maps feature to output channel dimension respectively. All 3 conv layers are designed with 3x3 kernels and stride 1.

Implementation Details. We implement Hierarchy Flow in the Pytorch framework, and train for 300k iterations using an Adam optimizer with a batch size of 1, an initial learning rate of 1e-5, and a cosine annealing scheduler which continuously decreases the learning rate to 0. The loss weight λ is set to 0.1, and k in aligned-style loss is set to 0.8 unless specified. In each experiment, we train the model with 10 random seeds and report the average quantitative results among them.

4.2 GTA to Cityscapes [Strong-Fidelity]

Dataset: We use GTA dataset [6] as the source domain and Cityscapes [7] as the target domain. By default, all images are resized to 512×256 and randomly cropped to 256×256 during training. Evaluation is performed in 512×256.

Qualitative Evaluation: For the comparison to the previous work, we select several generic I2I models that focus on better semantic alignment and a specific photo-realism model EPE [19] trained with auxiliary inputs. As shown in Figure 6, except VSAIT [34], all previous I2I models fail to retain full content information, and different levels of

TABLE 2
Quantitative results of art-style transfer. ArtFlow with $*$ and † represents the combination with AdaIN and WCT respectively. SSIM and KID($\times 10^3$) are used as metrics. Our model achieved competitive results on both stylization and content preservation with the lowest parameter number and FLOPs.

Method	SSIM↑	KID↓	Params↓	FLOPs↓
AdaIN [12]	0.28	41.1/5.1	7.01M	117.5G
WCT [13]	0.24	51.2/6.2	34.24M	272.3G
ArtFlow* [23]	0.52	**24.6/3.8**	6.42M	105.2G
ArtFlow† [23]	0.53	33.3/5.3	6.42M	105.2G
CCPL [36]	0.43	39.1/6.8	8.67M	90.2G
ours	**0.60**	28.2/4.7	**1.01M**	**24.6G**

TABLE 3
Human preference score. “Detail” and “Overall” denote the evaluation criteria of content preservation and overall performance. Our model surpass the previous SOTA methods CCPL [36] by **68.1%** in “Detail” and **30.3%** in “Overall”.

Method	**Ours**	CCPL [36]	ArtFlow [23]	AdaIN [12]	WCT [13]
Detail↑	**76.5%**	8.40%	15.1%	0%	0%
Overall↑	**47.9%**	17.6%	25.2%	5.9%	3.3%

TABLE 4
Quantitative results of low-light enhancement. NIQE [51] score is used as the metric, where the smaller value indicates better perceptual performance. Our method consistently yields better results.

Method	MEF↓	LIME↓	NPE↓	DICM↓	All↓
Source Image	4.265	4.438	4.319	4.255	4.134
RetinexNet [2]	4.149	4.420	4.485	4.200	3.920
CycleGAN [1]	3.782	3.276	4.036	3.560	3.554
LLNet [16]	4.845	4.940	4.78	4.809	4.751
Jinag *et al.* [52]	**3.232**	3.719	4.113	3.570	3.385
ArtFlow [23]	3.621	3.579	3.052	3.578	3.381
Ours	3.511	**3.418**	**3.460**	**2.916**	**3.306**

content distortion exist. More details are shown in the blue box and orange box in Figure 6).

Quantitative Evaluation: To quantitatively evaluate the performance of GTA to Cityscapes, it is critical to choose suitable metrics that can measure the ability of content preservation during translation. As illustrated in [21], popular metrics like FID and KID ignore semantic mismatch during evaluation and thus can be misleading. Instead, in GTA to Cityscapes translation, we can utilize the semantic correspondence between images and segmentation labels as a reference during evaluation. Specifically, for each method, we use a lightweight DeepLabV3 [47] model to train on translated GTA images and segmentation masks and report the semantic segmentation evaluation on the validation set of Cityscapes, which reflects the performance of both content preservation and stylization at the same time. Additionally, we report the Structural Similarity Index Measure (SSIM) between translated images and source images, which also measures the performance of content preservation. Since EPE uses additional “G-buffers” information which is not publicly available to reproduce their method, the quantitative result of EPE is omitted. As shown in Table 1, our model outperforms previous methods by a large margin including the current SOTA method VSAIT.

4.3 Artistic Style Transfer [Normal-Fidelity]

Dataset: Following previous artistic style transfer work, we use MS-COCO [48] as source domain and Wiki-Art [3] as target domain in our experiments. By default, all images are resized to 300 × 400 for training and testing.

Qualitative Evaluation: We compare the visual performance with different methods. WCT [13] generates stylized images with severe content distortion. AdaIN [12] preserves content information to a certain extent while detailed textures are lost. Artflow [23] uses a flow-based network to prevent content distortion, while checkerboard artifacts exist due to the squeeze operation. CCPL [36] utilizes a novel transformation in replacement of AdaIN [12] and achieves good stylistic results. As shown in Figure 7, our results not only have great stylistic effects but also achieve the best retention of content information.

Quantitative Evaluation: Following [49], we evaluate the stylized images quantitatively using SSIM and KID, where SSIM indicates the performance of content preservation, KID measures the similarity between the transferred image and the target image. As shown in Table 2, our model achieves the best content preservation and the second-best KID score with over 5 times smaller parameters and FLOPs.

User Study: To give an additional quantitative evaluation, we conduct a user study from 119 volunteers. We randomly choose 42 content images and 26 style images from the test set to generate 1092 content-style pairs for each method. Each participant is randomly allocated 10-15 pairs and chooses the best method in *Detail* (preservation of texture and semantic information) and in *Overall* (overall performance, i.e., quality, stylization, fidelity) for each pair. We finally collect 1673 effective votes, and Table 3 shows the human performance rate, where our model outperforms previous methods by a large margin.

4.4 Low-level Vision [Strong-Fidelity]

We evaluated our model with two different low-level vision tasks: (1) Low-light Enhancement (2) Dehazing.

Dataset: We conduct the low-light enhancement experiment following the official LOL dataset [2]’s train/val/test split. Dehazing experiment is trained on RESIDE-ITS [50] and evaluated on Synthetic Objective Testing Set (SOTS) [50]. By default, all images are resized to 512 × 512 for training and testing.

Low-light Enhancement: We performed a quantitative comparison of our model and other methods on the LOL testset [2] with metric NIQE [51]. As shown in Table 4, our model achieved the best NIQE score on natural images compared to previous methods. More visual results are shown in Fig. 8.

Dehazing: We compare our model with previous methods in Table 5 with metrics PSNR and SSIM, our model achieved second-best PSNR and SSIM score. Fig. 9 illustrates qualitative results of our method.

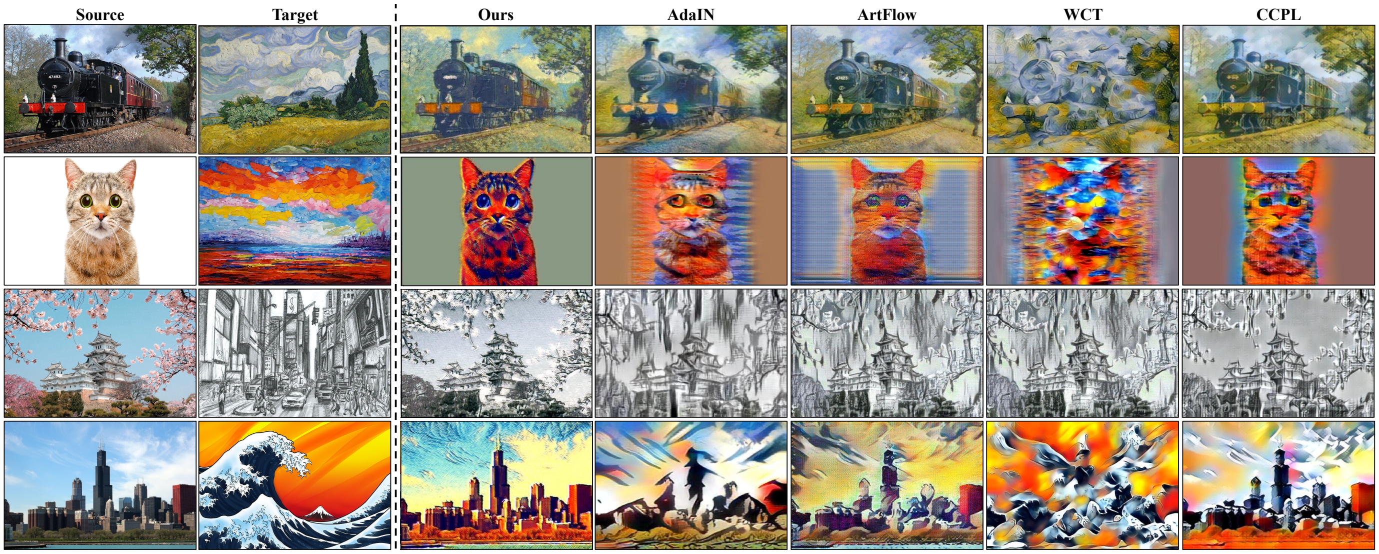

Fig. 7. Style transfer results compared with the state-of-the-art style transfer methods. Compared to other methods, our model generates images with satisfying artistic style without losing content information.

TABLE 5
Quantitative results of Dehaze. Our model achieves competitive results on both PSNR and SSIM metrics.

Method	DehazeNet [53]	GMAN [54]	GFN [55]	MAXIM [31]	GCANet [56]	ArtFlow [23]	**Ours**
PSNR↑	22.45	28.07	21.55	**34.09**	19.98	28.02	28.25
SSIM↑	0.851	0.934	0.843	**0.984**	0.704	0.935	0.945

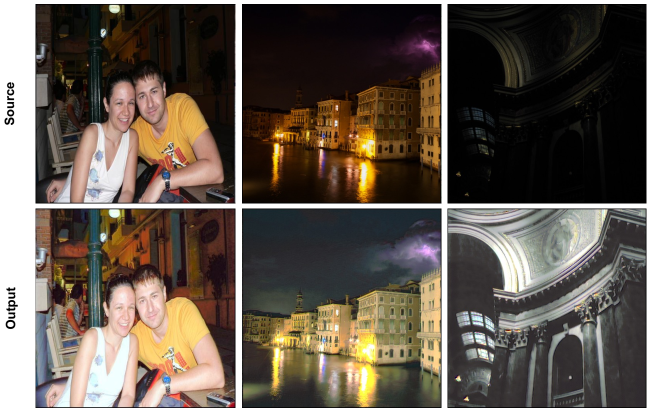

Fig. 8. Visual results of low-light enhancement on the LOL dataset [2]. Our generated images effectively retain content information and notably enhance dark areas compared to the source images.

4.5 Ablation Study

Runtime Analysis. On a single NVIDIA 32G V100 GPU, our method could process 86.46 samples per second in the GTA2Cityscapes task and 4.30 samples per second in the style-transfer task, which outperforms most SOTA methods in both tasks (see Table 6).

Architecture Analysis. To demonstrate the effectiveness of our network design, we conduct an ablation study on network architecture between our HF and ArtFlow [23], by adding the key components of ArtFlow except for the squeeze operation to our model. Table 8 shows the quantitative comparison of the artistic style transfer task. Compared to ArtFlow, our model achieves better performance with simpler network architecture, which shows the significance of Hierarchy Flow for I2I translation tasks.

Effect of k in *aligned-style loss*. Table 7 shows the effect of k in the *aligned-style loss* for the artistic style transfer task. As designed, larger k trades content preservation for stylization in translated images. The model trained with $k = 0.7$ retain the richest semantic details with the best SSIM score. Aligned-style loss with $k = 1.0$ is equivalent to the vanilla style loss and highly distorts semantic information and becomes over-stylized.

Effect of Model Size. We conduct ablation study of model size in Dehazing task. As shown in Table 7, base model *HF* with only 0.68M parameters already achieves competitive performance. As the model size expands (*HF*+: 0.74M; *HF*++: 1.01M; *HF*†: 6.30M), performance improves consistently.

Effect of Image Resolution. Due to the simplicity of network design, our model is capable to support high-dimensional image training and inference. We conduct an ablation study of different input/output resolutions with model *HF* and $k = 0.8$ for GTA to Cityscapes tasks. As results shown in Table 8, compared to the baseline in 512×256, mIoU obtains a performance boost by **+2.90** and **+4.28** with resolution of 1024×512 and 2048×1024 respectively.

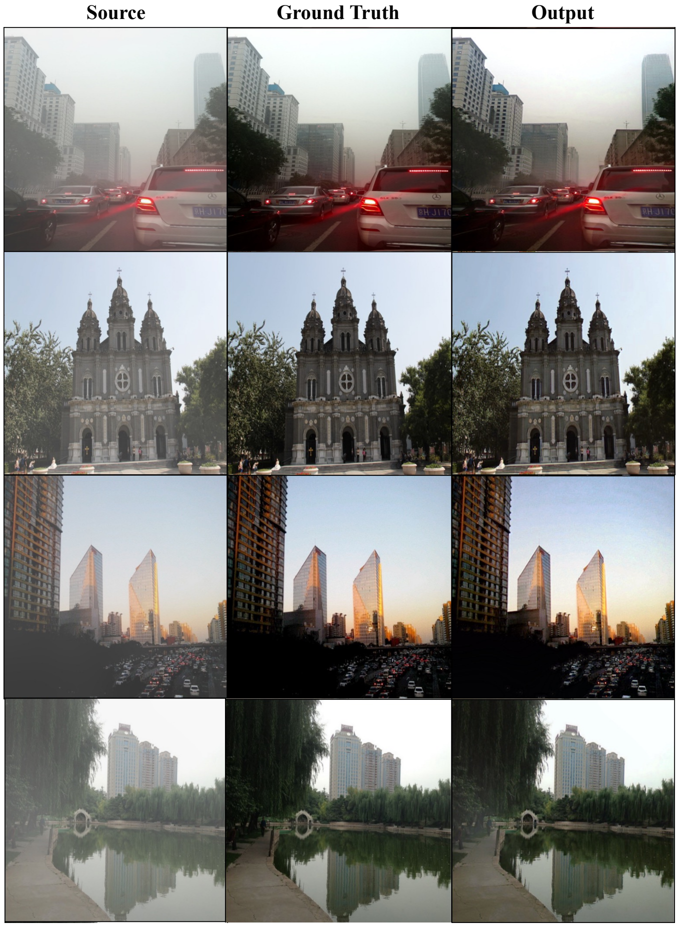

Fig. 9. Visual results of Dehaze. Our method shows outstanding dehazing performance in various scenarios, with output image quality comparable to the ground-truth images.

TABLE 6
Runtime analysis. Comparison on throughput (N samples per second) for different methods in GTA to Cityscapes (top rows) and artistic style transfer (bottom rows) respectively.

CycleGAN	GcGAN	DRIT++	CUT	VSAIT	**Ours**
15.76	33.15	21.06	32.39	32.39	**86.46**
CCPL	ArtFlow*	ArtFlow†	AdaIN	WCT	**Ours**
4.28	2.44	3.45	**14.66**	1.34	4.30

TABLE 7
Ablation study. 1. Ablation study of k in our proposed *aligned-style loss* on artistic style transfer task. Different k performs trade-off between content preservation and stylization, indicated by SSIM and FID respectively. 2. Model performance vs. Model size on Dehazing task, large model boosts performance consistently.

k	FID↓	SSIM↑	Model	PSRN↑	SSIM↑	Params↓
0.7	0.91	**0.60**	HF	26.76	0.931	**0.68M**
0.8	0.82	0.56	HF+	27.18	0.936	0.74M
0.9	**0.50**	0.39	HF++	27.45	0.937	1.01M
1.0	0.95	0.17	HF†	**28.25**	**0.945**	6.30M

5 Conclusion

In this paper, we categorize image-to-image translation problems into three levels: strong-, normal-, and weak-fidelity translation. We proposed a novel invertible network Hierarchy Flow, with a dedicated aligned-style loss for high-fidelity image-to-image translation. Qualitative and quantitative results show that our model obtains better content preservation during translation, and achieves the best performance in high-fidelity translation tasks.

TABLE 8
Ablation study of network architecture between HF and ArtFlow for artistic style transfer task. Overall, Hierarchy Flow achieves the best results.

Architecture	FID↓	SSIM↑	KID($\times 10^3$)↓
Hierachy Flow	**0.61**	**0.55**	**25.0/5.1**
+ Actnorm	0.88	0.29	27.3/4.9
+ 1x1 Conv	0.97	0.24	25.3/4.2
ArtFlow	0.85	0.21	27.2/4.5

TABLE 9
GTA to Cityscapes with different image resolutions. Due to the simple design of network, HF supports HD image training and testing, which yields better performance.

resolution	pxAcc↑	clsAcc↑	mIoU↑
512x256	81.04	41.92	31.52
1024x512	84.95	45.25	34.42
2048x1024	**87.21**	**46.74**	**35.80**

Future work. Although our model outperforms previous methods in strongly and normally constrained tasks, we failed to achieve admirable results in all weakly constrained translation tasks. Future work includes extending this model to full spectrum of image-to-image translation tasks.

Broader Impact. Our proposed generative model could eliminate the gap between simulation and reality, which can be widely used in self-driving and medical areas. The use of image synthesis would not lead to privacy issues but might create fake news, thus more regulations are needed to restrict the usage of synthesized data.

References

[1] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[2] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," *arXiv preprint arXiv:1808.04560*, 2018.

[3] K. Nichol, "Painter by numbers, wikiart," https://www.kaggle.com/c/painter-by-numbers, 2016.

[4] C. O. Ancuti, C. Ancuti, M. Sbert, and R. Timofte, "Dense haze: A benchmark for image dehazing with dense-haze and haze-free images," in *IEEE International Conference on Image Processing (ICIP)*, ser. IEEE ICIP 2019, 2019.

[5] C. O. Ancuti, C. Ancuti, R. Timofte, L. V. Gool, L. Zhang, and M.-H. Yang, "Ntire 2019 image dehazing challenge report," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, ser. IEEE CVPR 2019, 2019.

[6] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *European conference on computer vision*. Springer, 2016, pp. 102–118.

[7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[8] T. Li, R. Qian, C. Dong, S. Liu, Q. Yan, W. Zhu, and L. Lin, "Beautygan: Instance-level facial makeup transfer with deep generative adversarial network," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 645–653.

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[10] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," *Advances in neural information processing systems*, vol. 28, pp. 262–270, 2015.

[11] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[12] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.

[13] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," *arXiv preprint arXiv:1705.08086*, 2017.

[14] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[15] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[16] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017.

[17] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[18] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[19] S. R. Richter, H. A. AlHaija, and V. Koltun, "Enhancing photorealism enhancement," *arXiv preprint arXiv:2105.04619*, 2021.

[20] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Multimodal image-to-image translation by enforcing bi-cycle consistency," in *Advances in neural information processing systems*, 2017, pp. 465–476.

[21] Z. Jia, B. Yuan, K. Wang, H. Wu, D. Clifford, Z. Yuan, and H. Su, "Semantically robust unpaired image translation for data with unmatched semantics statistics," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 273–14 283.

[22] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. PMLR, 2018, pp. 1989–1998.

[23] J. An, S. Huang, Y. Song, D. Dou, W. Liu, and J. Luo, "Artflow: Unbiased image style transfer via reversible neural flows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 862–871.

[24] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," *arXiv preprint arXiv:1605.08803*, 2016.

[25] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *arXiv preprint arXiv:1807.03039*, 2018.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[27] L. Ma, X. Jia, S. Georgoulis, T. Tuytelaars, and L. Van Gool, "Exemplar guided unsupervised image-to-image translation with semantic consistency," *arXiv preprint arXiv:1805.11145*, 2018.

[28] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 35–51.

[29] M. Kumar, D. Weissenborn, and N. Kalchbrenner, "Colorization transformer," *arXiv preprint arXiv:2102.04432*, 2021.

[30] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu, "Attentive generative adversarial network for raindrop removal from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2482–2491.

[31] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxim: Multi-axis mlp for image processing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5769–5780.

[32] D. Engin, A. Genç, and H. Kemal Ekenel, "Cycle-dehaze: Enhanced cyclegan for single image dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 825–833.

[33] Z. Anvari and V. Athitsos, "Dehaze-glcgan: unpaired single image de-hazing via adversarial training," *arXiv preprint arXiv:2008.06632*, 2020.

[34] J. Theiss, J. Leverett, D. Kim, and A. Prakash, "Unpaired image translation via vector symbolic architectures," in *European Conference on Computer Vision*. Springer, 2022, pp. 17–32.

[35] T. Q. Chen and M. Schmidt, "Fast patch-based style transfer of arbitrary style," *arXiv preprint arXiv:1612.04337*, 2016.

[36] Z. Wu, Z. Zhu, J. Du, and X. Bai, "Ccpl: Contrastive coherence preserving loss for versatile style transfer," in *European Conference on Computer Vision*. Springer, 2022, pp. 189–206.

[37] X. Li, K. Kou, and B. Zhao, "Weather gan: Multi-domain weather translation using generative adversarial networks," *arXiv preprint arXiv:2103.05422*, 2021.

[38] J. Lin, Y. Xia, T. Qin, Z. Chen, and T.-Y. Liu, "Conditional image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5524–5532.

[39] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 172–189.

[40] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *European Conference on Computer Vision*. Springer, 2020, pp. 319–345.

[41] I. Kobyzev, S. Prince, and M. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[42] L. Dinh, D. Krueger, and Y. Bengio, "Nice: Non-linear independent components estimation," *arXiv preprint arXiv:1410.8516*, 2014.

[43] J. Ho, X. Chen, A. Srinivas, Y. Duan, and P. Abbeel, "Flow++: Improving flow-based generative models with variational dequantization and architecture design," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2722–2730.

[44] H.-J. Chen, K.-M. Hui, S.-Y. Wang, L.-W. Tsao, H.-H. Shuai, and W.-H. Cheng, "Beautyglow: On-demand makeup transfer framework with reversible generative network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 042–10 050.

[45] H. Fu, M. Gong, C. Wang, K. Batmanghelich, K. Zhang, and D. Tao, "Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2427–2436.

[46] H.-Y. Lee, H.-Y. Tseng, Q. Mao, J.-B. Huang, Y.-D. Lu, M. Singh, and M.-H. Yang, "Drit++: Diverse image-to-image translation via disentangled representations," *International Journal of Computer Vision*, vol. 128, no. 10, pp. 2402–2417, 2020.

[47] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[48] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[49] K. Hong, S. Jeon, H. Yang, J. Fu, and H. Byun, "Domain-aware universal style transfer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 14 609–14 617.

[50] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2018.

[51] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[52] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "Enlightengan: Deep light enhancement

without paired supervision," *IEEE Transactions on Image Processing*, vol. 30, pp. 2340–2349, 2021.

[53] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5187–5198, 2016.

[54] Z. Liu, B. Xiao, M. Alrabeiah, K. Wang, and J. Chen, "Single image dehazing with a generic model-agnostic convolutional neural network," *IEEE Signal Processing Letters*, vol. 26, no. 6, pp. 833–837, 2019.

[55] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, and M.-H. Yang, "Gated fusion network for single image dehazing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3253–3261.

[56] D. Chen, M. He, Q. Fan, J. Liao, L. Zhang, D. Hou, L. Yuan, and G. Hua, "Gated context aggregation network for image dehazing and deraining," in *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2019, pp. 1375–1383.

Weichen Fan received his B.S. degree in electronic science and engineering (ESE) from the University of Electronic Science and Technology of China (UESTC) and his master's degree from the Department of Electronic and Computer Engineering (ECE), National University of Singapore (NUS). His is currently a researcher at SenseTime. His research interests include low-level vision, transfer learning, and multi-modal learning.

Jinghuan Chen received his B.Eng. degree from Nanyang Technological University, Singapore in 2020. He is currently an Algorithm Engineer at ByteDance Inc. His research interests include computer vision, generative models and multi-modal learning.

Ziwei Liu is currently a Nanyang Assistant Professor at Nanyang Technological University, Singapore. His research revolves around computer vision, machine learning and computer graphics. He has published extensively on top-tier conferences and journals in relevant fields, including CVPR, ICCV, ECCV, NeurIPS, ICLR, ICML, TPAMI, TOG and Nature - Machine Intelligence. He is the recipient of Microsoft Young Fellowship, Hong Kong PhD Fellowship, ICCV Young Researcher Award, HKSTP Best Paper Award and WAIC Yunfan Award. He serves as an Area Chair of CVPR, ICCV, NeurIPS and ICLR, as well as an Associate Editor of IJCV.