TSIT: A Simple and Versatile Framework for Image-to-Image Translation

Supplementary Material

Liming Jiang[1], Changxu Zhang[2], Mingyang Huang[3], Chunxiao Liu[3], Jianping Shi[3], and Chen Change Loy[1]✉

[1] Nanyang Technological University, Singapore, Singapore
{liming002,ccloy}@ntu.edu.sg
[2] University of California, Berkeley, CA, USA
zhangcx@berkeley.edu
[3] SenseTime Research, Beijing, China
{huangmingyang,liuchunxiao,shijianping}@sensetime.com

Abstract. This document provides the supplementary information that is not elaborated in our main paper due to the space constraints: Sec. A details the different applications we have explored. Sec. B describes details of the datasets used in our experiments. Sec. C provides additional implementation details. Sec. D presents several supplementary ablation studies. Sec. E shows more examples of the images generated by our method.

A Application Exploration

We have introduced a Two-Stream Image-to-image Translation (TSIT) framework in the main paper. The proposed framework is simple and versatile for various image-to-image translation tasks under both unsupervised and supervised settings. We have considered three important and representative applications of conditional image synthesis: arbitrary style transfer (unsupervised), semantic image synthesis (supervised), and multi-modal image synthesis (enriching generation diversity). We employ a two-stream network, namely "content" stream and "style" stream, on these applications.

For the unsupervised arbitrary style transfer application, we feed the content image to the content stream and the style image to the style stream, and let the networks learn different levels of *feature representations* of the content and style. The proposed feature transformations, FADE and FAdaIN, adaptively fuse content and style feature maps, respectively, at different scales in the generator. In contrast to prior works, our method is more adaptable to style transfer tasks in diverse scenarios (*e.g.*, natural images, real-world scenes, and artistic paintings).

We further expand the application of our method to cater to semantic image synthesis under the supervised setting. The definition of "content" and "style" can be more general: all the images that provide semantic structure information

can be content images, and all the images representing the global style distribution can be considered as style images. Therefore, when we inject semantic segmentation masks to the content stream and the corresponding real images to the style stream, semantic image synthesis task in the supervised setting can be handled. Despite a rather large domain gap in this task, our framework yields comparable or even better results over the state-of-the-art task-specific methods, suggesting the high adaptability of our approach.

It is noteworthy to highlight that the newly proposed feature transformations and the symmetrical two-stream network can effectively disentangle the semantic structure and style information. Thanks to the clean disentanglement, the high-level multi-modal nature of the images can be captured by our framework, contributing to high-fidelity multi-modal image synthesis.

B Dataset Details

In this section, we discuss the detailed information of all the datasets we explored, including the source, preprocessing, number of images, resolution, *etc.*

For arbitrary style transfer under the unsupervised setting, paired data are not needed. We perform style transfer tasks in diverse scenarios (*e.g.*, natural images, real-world scenes, and artistic paintings).

- **Yosemite summer → winter.** We use this unpaired dataset provided by [12], containing rich natural images collected via Flickr API. We perform season transfer using this dataset, with $1,231$ summer images and 962 winter images for training. The resolution is 256×256.
- **BDD100K day → night.** We conduct time translation on BDD100K [10] dataset, which is captured at diverse locations in the United States. All the images are in real-world scenes, mostly street/road scenes. We classify the dataset into different times. The training set contains $12,454$ daytime images and $22,884$ nighttime images. The original images are scaled to 512×256.
- **Photo → art.** We utilize the art dataset collected in [12]. The art images of this dataset were downloaded from Wikiart.org. The dataset consists of photographs and diverse artistic paintings (Monet: $1,074$; Cézanne: 584; van Gogh: 401; Ukiyo-e: $1,433$). To test the robustness of the models for arbitrary style transfer, we combine all the artistic styles, yielding $6,853$ photos and $3,492$ paintings for training. All the images are uniformly resized to 256×256.

For semantic image synthesis under the supervised setting, we follow [6, 7] and select several challenging datasets.

- **Cityscapes.** Cityscapes [1] dataset contains street scene images mostly collected in Germany, with $2,975$ images for training and 500 images for evaluation. The dataset provides instance-wise, dense pixel annotations of 30 classes. All the image sizes are adjusted to 512×256.
- **ADE20K.** We use ADE20K [11] dataset consisting of challenging in-the-wild images with fine annotations of 150 semantic classes. The sizes of training and validation sets are $20,210$ and $2,000$, respectively. All the images are scaled to 256×256.

For multi-modal image synthesis, we use BDD100K [10] dataset, details of which have been described earlier.

– **BDD100K sunny → different time/weather conditions**. We further classify the images in BDD100K [10] dataset into different time and weather conditions, constituting a training set of $10,000$ sunny images and $10,000$ images of other time and weather conditions (night: $2,500$; cloudy: $2,500$; rainy: $2,500$; snowy: $2,500$). The resolution is 512×256.

C Additional Implementation Details

We provide more implementation details in this section, including the network architecture specifics, detailed feature shapes, hyperparameters, *etc.*

Network architecture specifics. Our framework consists of four components: content stream, style stream, generator, and discriminators. The first three components maintain a symmetrical structure, using fully convolutional networks. The number of residual blocks k (*i.e.*, downsampling/upsampling times) in the content/style stream and the generator equals to 7. Let inc, $outc$, kn, s, p denote the input channel, the output channel, the kernel size, the stride, and the zero-padding amount, respectively.

In the content/style stream, we use a series of content/style residual blocks with the nearest neighbor downsampling. The scale factor of downsampling is 2. By default, we use instance normalization [8] for the content/style residual blocks, and the negative slope of Leaky ReLU is 0.2. Thus, the structure of Content/Style ResBlk($inc, outc$) is: Downsample(2)$-$Conv($inc, inc, kn3\times 3, s1, p1$)$-$IN $-$ LReLU(0.2)$-$Conv($inc, outc, kn3 \times 3, s1, p1$)$-$IN $-$ LReLU(0.2) with the learned skip connection Conv($inc, outc, kn1\times 1, s1, p0$)$-$IN $-$ LReLU(0.2).

In the generator, we construct several FADE residual blocks with the nearest neighbor upsampling. The scale factor of upsampling is 2. FAdaIN layers are applied before each FADE residual block. The FADE residual block contains a FADE submodule, which performs *element-wise* denormalization using a learned affine transformation defined by the modulation parameters γ and β. Let $normc$, $featc$ indicate the normalized channel and the injected feature channel, respectively. Then, the convolutional layers in FADE($normc, featc$) can be represented as: Conv($featc, normc, kn3 \times 3, s1, p1$). By default, we adopt SyncBN for the generator, and the negative slope of Leaky ReLU is 0.2. The structure of FADE ResBlk($inc, outc$) is: FADE(inc, inc)$-$LReLU(0.2)$-$Conv($inc, inc, kn3 \times 3, s1, p1$)$-$FADE($inc, inc$)$-$LReLU(0.2)$-$Conv($inc, outc, kn3 \times 3, s1, p1$)$-$Upsample(2) with the learned skip connection FADE(inc, inc)$-$LReLU(0.2)$-$Conv($inc, outc, kn1 \times 1, s1, p0$).

As mentioned in the main paper, we exploit the same multi-scale patch-based discriminators as [9, 7]. The detailed network architectures and the layers used for feature matching loss [9] are also identical.

Feature shapes. In the content/style stream, we put an input layer at the entrance. The feature channel is adjusted to 64 after the input layer, while the resolution remains unchanged. Then, the feature channels after each of the $k(7)$

residual blocks are: 128, 256, 512, 1024, 1024, 1024, 1024. Since the scale factor of downsampling is 2 (as described in the network architecture specifics above), the resolution of the features is halved after each residual block. The generator feature shapes are strictly corresponding and opposite to that of content/style stream. The discriminator feature shapes are identical to that in [9, 7], where the resolution is halved on every step of the pyramid.

Additional training details. For perceptual loss, we use the feature reconstruction loss that requires a content target [4].

In the arbitrary style transfer and multi-modal image synthesis tasks, the content target is the content image. The loss weights are $\lambda_P = 1, \lambda_{FM} = 1$, and the batch size is 1. We train our models for 200 epochs on Yosemite summer $\rightarrow$ winter, 10 epochs on BDD100K day $\rightarrow$ night, 40 epochs on Photo $\rightarrow$ art, and 20 epochs on BDD100K sunny $\rightarrow$ different time/weather conditions. The models are trained on 1 NVIDIA Tesla V100 GPU, with around 10 GB memory consumption. For multi-modal image synthesis, similar to [3], at the inference phase we run the generator network in exactly the same manner as during the training phase. For the cross validation of SPADE [7], the hyperparameters obtaining the best generation results are $\lambda_P = 10, \lambda_{FM} = 10$.

In the semantic image synthesis task, the content target is the ground truth real image. The corresponding loss weights are $\lambda_P = 20, \lambda_{FM} = 10$, and the batch size is 16. We perform 200 epochs of training on Cityscapes and ADE20K. The models are trained on 2 NVIDIA Tesla V100 GPUs, each with about 32 GB memory consumption. We also find that in semantic image synthesis, weakening/removing the style stream can sometimes contribute to a performance boost. Besides, exploiting variational auto-encoders [5] can help in certain cases. For the cross validation of MUNIT [2], since the loss functions are very different from ours, we use its default hyperparameters in unsupervised image-to-image translation.

D Supplementary Ablation Studies

We ablate the key modules (*i.e.*, content stream (CS), style stream(SS)) and the proposed feature transformations in the main paper. We perform multi-modal image synthesis to clearly show the effectiveness of different components. Due to the space constraints, we only provide qualitative evaluation results. In this section, we will first show the quantitative evaluation results of key component ablation studies in the main paper. Then, we will dig deeper and present more supplementary ablation study results.

Quantitative evaluation of key component ablation studies. We conduct quantitative evaluation on ablation studies of the key components in multi-modal image synthesis task. As shown in Table 1, using the full model we introduced, the lowest FID score and highest IS score have been achieved. This means the generated images by our full model are the most photorealistic, clearest, and of the highest diversity. Without any key module of TSIT, the quantitative

Table 1. The quantitative evaluation on ablation studies of the key modules (*i.e.*, content stream (CS), style stream (SS)) and the feature transformations in multi-modal image synthesis task. A lower FID and a higher IS indicate better performance.

	multi-modal image synthesis				
Metrics	full model	w/o CS	w/o SS	w/o FADE	w/o FAdaIN
FID ↓	**85.876**	89.429	86.263	86.463	89.795
IS ↑	**2.934**	2.851	2.734	2.881	2.890

Table 2. The quantitative evaluation on ablation studies of CS/SS feature channels for unsupervised arbitrary style transfer (day → night). A lower FID and a higher IS indicate better performance.

	arbitrary style transfer (day → night)		
Metrics	full model	channels ÷ 2	channels ÷ 4
FID ↓	**79.697**	82.357	95.199
IS ↑	**2.203**	2.142	2.101

Table 3. The quantitative evaluation on ablation studies of CS/SS feature channels for supervised semantic image synthesis (Cityscapes). A higher mIoU, a higher pixel accuracy (accu) and a lower FID indicate better performance.

	semantic image synthesis (Cityscapes)		
Metrics	full model	channels ÷ 2	channels ÷ 4
mIoU ↑	**65.9**	61.0	56.6
accu ↑	**94.4**	93.7	93.0
FID ↓	**59.2**	71.8	74.4

Table 4. The quantitative evaluation on ablation studies of CS/SS feature channels for multi-modal image synthesis. A lower FID and a higher IS indicate better performance.

	multi-modal image synthesis		
Metrics	full model	channels ÷ 2	channels ÷ 4
FID ↓	**85.876**	93.258	97.297
IS ↑	**2.934**	2.851	2.813

performance will drop. This verifies the necessity of these components for our method.

Feature channel ablation studies. We also study how the number of feature channels in the two streams (*i.e.*, content stream (CS) and style stream (SS)) affects the image synthesis results. We conduct quantitative evaluation of feature channel ablation studies, covering all of the discussed tasks. Note that we should change the channels in CS/SS at the same time to maintain a symmetrical structure. As presented in Table 2, Table 3 and Table 4, in different tasks under either unsupervised or supervised setting, the best performance is achieved by the full model of TSIT. As we reduce the channel numbers in the

Table 5. The quantitative evaluation on ablation studies of feature-level (FAdaIN)/ image-level (AdaIN) injection for unsupervised arbitrary style transfer (day → night). A lower FID and a higher IS indicate better performance.

	arbitrary style transfer (day → night)	
Metrics	feature-level	image-level
FID ↓	**79.697**	80.618
IS ↑	**2.203**	2.182

Table 6. The quantitative evaluation on ablation studies of feature-level (FADE)/ image-level (SPADE) injection for supervised semantic image synthesis (Cityscapes). A higher mIoU, a higher pixel accuracy (accu) and a lower FID indicate better performance.

	semantic image synthesis (Cityscapes)	
Metrics	feature-level	image-level
mIoU ↑	**65.9**	59.7
accu ↑	**94.4**	93.3
FID ↓	**59.2**	60.1

two-stream network, the image synthesis quality gradually degrade. For more channels, memory consumption will increase exponentially.

Feature-level/Image-level injection ablation studies. To verify the importance of the feature-level injection, We further conduct feature-level/image-level injection ablation studies. TSIT performs feature-level injections from the content/style stream to the generator to adapt to diverse tasks. In comparison, the direct injection of resized images (*i.e.*, the direct application of AdaIN in arbitrary style transfer, and SPADE in semantic image synthesis) can be regarded as the image-level injections. We provide quantitative evaluation results under this setting. As shown in Table 5 and Table 6, compared to our feature-level injection scheme, the image-level injection leads to a performance drop. This suggests the significance of feature-level injection in TSIT.

E More Examples of Generated Images

We show more examples of generated results by our method in Fig. 1 and Fig. 2. Several generated images of arbitrary style transfer, covering diverse scenarios, are presented in Fig. 1. We also show more synthesized exmaples of semantic image synthesis in Fig. 2. These examples feature both outdoor and indoor scenes, generated from the corresponding semantic segmentation label maps. All the images synthesized by our proposed method are very photorealistic.

References

1. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes dataset for semantic urban scene

understanding. In: CVPR (2016)
2. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV (2018)
3. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
4. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016)
5. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint **arXiv:1312.6114** (2013)
6. Liu, X., Yin, G., Shao, J., Wang, X., et al.: Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In: NeurIPS (2019)
7. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019)
8. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint **arXiv:1607.08022** (2016)
9. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: CVPR (2018)
10. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T.: BDD100K: A diverse driving video database with scalable annotation tooling. arXiv preprint **arXiv:1805.04687** (2018)
11. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ADE20K dataset. In: CVPR (2017)
12. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)

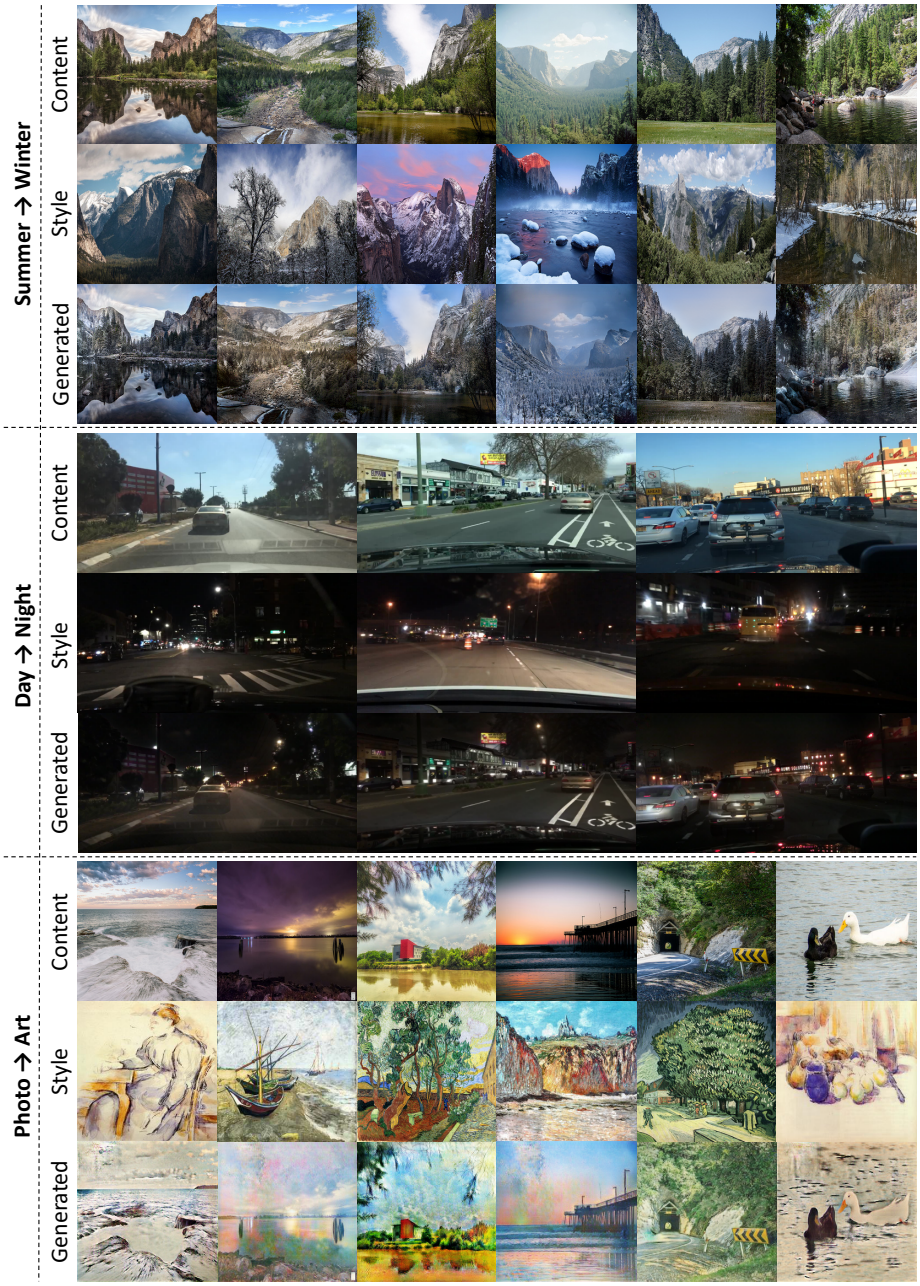

Fig. 1. More examples of images generated by our method in the arbitrary style transfer task (unsupervised). Rows 1-3 show Yosemite summer → winter season transfer results. Rows 4-6 are BDD100K day → night translation results. Rows 7-9 present photo → art style transfer results.

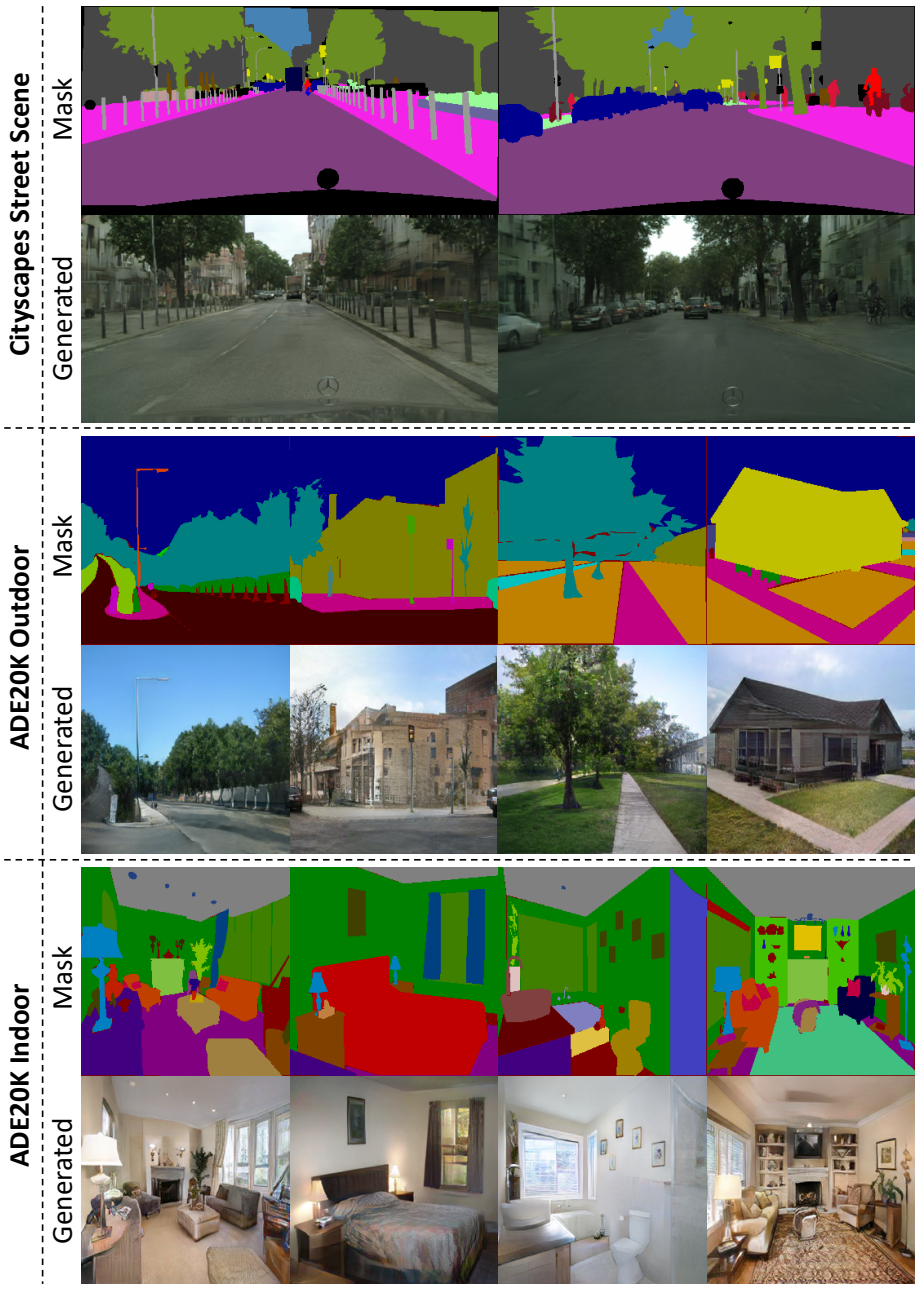

Fig. 2. More examples of images generated by our method in the semantic image synthesis task (supervised). Row 1 and 2 show generated results on Cityscapes dataset. Row 3 and 4 are outdoor synthesized results on ADE20K dataset. Row 5 and 6 present indoor synthesized results on ADE20K dataset.