Unsupervised Image-to-Image Translation with Stacked Cycle-Consistent Adversarial Networks

Minjun Li[1,2], Haozhi Huang[2], Lin Ma[2], Wei Liu[2], Tong Zhang[2], Yu-Gang Jiang[1]*

[1] Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University
[2] Tencent AI Lab
me@minjun.li, {huanghz08, forest.linma}@gmail.com,
wl2223@columbia.edu, tongzhang@tongzhang-ml.org, ygj@fudan.edu.cn

Abstract. Recent studies on unsupervised image-to-image translation have made a remarkable progress by training a pair of generative adversarial networks with a cycle-consistent loss. However, such unsupervised methods may generate inferior results when the image resolution is high or the two image domains are of significant appearance differences, such as the translations between semantic layouts and natural images in the Cityscapes dataset. In this paper, we propose novel Stacked Cycle-Consistent Adversarial Networks (SCANs) by decomposing a single translation into multi-stage transformations, which not only boost the image translation quality but also enable higher resolution image-to-image translations in a coarse-to-fine manner. Moreover, to properly exploit the information from the previous stage, an adaptive fusion block is devised to learn a dynamic integration of the current stage's output and the previous stage's output. Experiments on multiple datasets demonstrate that our proposed approach can improve the translation quality compared with previous single-stage unsupervised methods.

Keywords: Image-to-Image Translation · Unsupervised Learning · Generative Adversarial Network (GAN)

1 Introduction

Image-to-image translation attempts to convert the image appearance from one domain to another while preserving the intrinsic image content. Many computer vision tasks can be formalized as a certain image-to-image translation problem, such as super-resolution [14,20], image colorization [30,31,6], image segmentation [17,4], and image synthesis [1,21,26,13,33]. However, conventional image-to-image translation methods are all task specific. A common framework for universal image-to-image translation remains as an emerging research subject in the literature, which has gained considerable attention in recent studies [7,34,10,16,27].
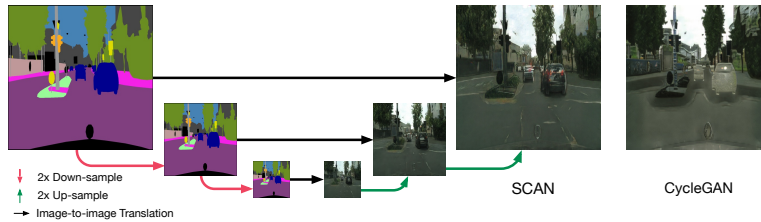
---

* The corresponding author.

Fig. 1. Given unpaired images from two domains, our proposed SCAN learns the image-to-image translation by a stacked structure in a coarse-to-fine manner. For the Cityscapes *Labels → Photo* task in 512×512 resolution, the result of SCAN (left) appears more realistic and includes finer details compared with the result of Cycle-GAN [34] (right).

Isola *et al.* [7] leveraged the power of generative adversarial networks (GANs) [5,18,32], which encourage the translation results to be indistinguishable from real images in the target domain, to learn image-to-image translation from image pairs in a supervised fashion. However, obtaining pairwise training data is time-consuming and heavily relies on human labor. Recent works [34,10,16,27] explore tackling the image-to-image translation problem without using pairwise data. Under the unsupervised setting, besides the traditional adversarial loss used in supervised image-to-image translation, a cycle-consistent loss is introduced to restrain the two cross-domain transformations G and F to be the inverses of each other (*i.e.*, $G(F(x)) \approx x$ and $G(F(y)) \approx y$). By constraining both of the adversarial and cycle-consistent losses, the networks learn how to accomplish cross-domain transformations without using pairwise training data.

Despite the progress mentioned above, existing unsupervised image-to-image translation methods may generate inferior results when two image domains are of significant appearance differences or the image resolution is high. As shown in Figure 1, the result of CycleGAN [34] in translating a Cityscapes semantic layout to a realistic picture lacks details and remains visually unsatisfactory. The reason for this phenomenon lies in the significant visual gap between the two distinct image domains, which makes the cross-domain transformation too complicated to be learned by running a single-stage unsupervised approach.

Jumping out of the scope of unsupervised image-to-image translation, many methods have leveraged the power of multi-stage refinements to tackle image generation from latent vectors [3,9], caption-to-image [29], and supervised image-to-image translation [1,4,23]. By generating an image in a coarse-to-fine manner, a complicated transformation is broken down into easy-to-solve pieces. Wang *et al.* [23] successfully tackled the high-resolution image-to-image translation problem in such a coarse-to-fine manner with multi-scale discriminators. However, their method relies on pairwise training images, so cannot be directly applied to our studied unsupervised image-to-image translation task. To the best of our

knowledge, there exists no attempt to exploit stacked networks to overcome the difficulties encountered in learning unsupervised image-to-image translation.

In this paper, we propose the stacked cycle-consistent adversarial networks (SCANs) aiming for unsupervised learning of image-to-image translation. We decompose a complex image translation into multi-stage transformations, including a coarse translation followed by multiple refinement processes. The coarse translation learns to sketch a primary result in low-resolution. The refinement processes improve the translation by adding details into the previous results to produce higher resolution outputs. We adopt a conjunction of an adversarial loss and a cycle-consistent loss in all stages to learn translations from unpaired image data. To benefit more from multi-stage learning, we also introduce an adaptive fusion block in the refinement processes to learn the dynamic integration of the current stage's output and the previous stage's output. Extensive experiments demonstrate that our proposed model can not only generate results with realistic details, but also enable us to learn unsupervised image-to-image translation in higher resolution.

In summary, our contributions are mainly two-fold. Firstly, we propose SCANs to model the unsupervised image-to-image translation problem in a coarse-to-fine manner for generating results with finer details in higher resolution. Secondly, we introduce a novel adaptive fusion block to dynamically integrate the current stage's output and the previous stage's output, which outperforms directly stacking multiple stages.

2   Related Work

Image-to-image translation. GANs [5] have shown impressive results in a wide range of tasks including super-resolution [14,20], video generation [28], image colorization [7], image style transfer [34] etc. The essential part of GANs is the idea of using an adversarial loss that encourages the translated results to be indistinguishable from real target images. Among the existing image-to-image translation works using GANs, perhaps the most well-known one would be Pix2Pix [7], in which Isola *et al.* applied GANs with a regression loss to learn pairwise image-to-image translation. Due to the fact that pairwise image data is difficult to obtain, image-to-image translation using unpaired data has drawn rising attention in recent studies. Recent works by Zhu *et al.* [34], Yi *et al.* [27], and Kim *et al.* [10] have tackled the image translation problem using a combination of adversarial and cycle-consistent losses. Taigman *et al.* [22] applied cycle-consistency in the feature level with the adversarial loss to learn a one-side translation from unpaired images. Liu *et al.* [16] used a GAN combined with Variational Auto Encoder (VAE) to learn a shared latent space of two given image domains. Liang *et al.* [15] combined the ideas of adversarial and contrastive losses, using a contrastive GAN with cycle-consistency to learn the semantic transform of two given image domains with labels. Instead of trying to translate one image to another domain directly, our proposed approach focuses on explor-

ing refining processes in multiple steps to generate a more realistic output with finer details by harnessing unpaired image data.

Multi-stage learning. Extensive works have proposed to choose multiple stages to tackle complex generation or transformation problems. Eigen *et al.* [4] proposed a multi-scale network to predict depth, surface, and segmentation, which learns to refine the prediction result from coarse to fine. S2GAN introduced by Wang *et al.* [24] utilizes two networks arranged sequentially to first generate a structure image and then transform it into a natural scene. Zhang *et al.* [29] proposed StackGAN to generate high-resolution images from texts, which consists of two stages: the Stage-I network generates a coarse, low-resolution result, while the Stage-II network refines the result into a high-resolution, realistic image. Chen *et al.* [1] applied a stacked refinement network to generate scenes from segmentation layouts. To accomplish generating high-resolution images from latent vectors, Kerras *et al.* [9] started from generating a 4×4 resolution output, and then progressively stacked up both a generator and a discriminator to generate a 1024×1024 realistic image. Wang *et al.* [23] applied a coarse-to-fine generator with a multi-scale discriminator to tackle the supervised image-to-image translation problem. Different form the existing works, this work exploits stacked image-to-image translation networks coupled with a novel adaptive fusion block to tackle the unsupervised image-to-image translation problem.

3   Proposed Approach

3.1   Formulation

Given two image domains X and Y, the mutual translations between them can be denoted as two mappings $G : X \rightarrow Y$ and $F : Y \rightarrow X$, each of which takes an image from one domain and translates it to the corresponding representation in the other domain. Existing unsupervised image-to-image translation approaches [34,27,10,16,22] finish the learning of G and F in a single stage, which generate results lacking details and are unable to handle complex translations.

In this paper, we decompose translations G and F into multi-stage mappings. For simplicity, now we describe our method in a two-stage setting. Specifically, we decompose $G = G_2 \circ G_1$ and $F = F_2 \circ F_1$. G_1 and F_1 (**Stage-1**) perform the cross-domain translation in a coarse scale, while G_2 and F_2 (**Stage-2**) serve as refinements on the top of the outputs from the previous stage. We first finish the training of Stage-1 in low-resolution and then train Stage-2 to learn refinement in higher resolution based on the fixed Stage-1.

Training two stages in the same resolution would make Stage-2 difficult to bring further improvement, as Stage-1 has already been optimized with the same objective function (see Section 4.5). On the other hand, we find that learning in a lower resolution allows the model to generate visually more natural results, since the manifold underlying the low-resolution images is easier to model. Therefore, first, we constrain Stage-1 to train on 2x down-sampled image samples, denoted by $X_\downarrow$ and $Y_\downarrow$, to learn a base transformation. Second, based on the outputs of
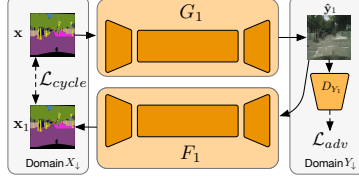
Fig. 2. Illustration of an overview of Stage-1 for learning coarse translations in low-resolution under an unsupervised setting. Solid arrow denotes an input-output, and dashed arrow denotes a loss.

Stage-1, we train Stage-2 with image samples X and Y in the original resolution. Such a formulation exploits the preliminary low-resolution results of Stage-1 and guides Stage-2 to focus on up-sampling and adding finer details, thus helping improve the overall translation quality.

In summary, to learn cross-domain translations $G : X \to Y$ and $F : Y \to X$ on given domains X and Y, we first learn preliminary translations $G_1 : X_\downarrow \to Y_\downarrow$ and $F_1 : Y_\downarrow \to X_\downarrow$ at the 2x down-sampled scale. Then we use $G_2 : X_\downarrow \to X$ and $F_2 : Y_\downarrow \to Y$ to obtain the final output with finer details in the original resolution. Notice that we can iteratively decompose G_2 and F_2 into more stages.

3.2   Stage-1: Basic Translation

In general, our Stage-1 module adopts a similar architecture of CycleGAN [34], which consists of two image translation networks G_1, and F_1 and two discriminators D_{X_1}, D_{Y_1}. Note that Stage-1 is trained in low-resolution image domains $X_\downarrow$ and $Y_\downarrow$. Figure 2 shows an overview of the Stage-1 architecture.

Given a sample $\mathbf{x}_1 \in X_\downarrow$, G_1 translates it to a sample $\hat{\mathbf{y}}_1 = G_1(\mathbf{x}_1)$ in the other domain $Y_\downarrow$. On one hand, the discriminator D_{Y_1} learns to classify the generated sample $\hat{\mathbf{y}}_1$ to class 0 and the real image $\mathbf{y}$ to class 1, respectively. On the other hand, G_1 learns to deceive D_{Y_1} by generating more and more realistic samples. This can be formulated as an adversarial loss:

$$\begin{aligned}
\mathcal{L}_{adv}(G_1, D_{Y_1}, X_\downarrow, Y_\downarrow) = &\, \mathbb{E}_{\mathbf{y} \sim Y\downarrow}\left[\log(D_{Y_1}(\mathbf{y}))\right] \\
&+ \mathbb{E}_{\mathbf{x} \sim X\downarrow}\left[\log(1 - D_{Y_1}(G_1(\mathbf{x})))\right].
\end{aligned} \tag{1}$$

While D_{Y_1} tries to maximize $\mathcal{L}_{adv}$, G_1 tries to minimize it. Afterward, we use F_1 to translate $\hat{\mathbf{y}}_1$ back to the domain $X_\downarrow$, and constrain $F_1(\hat{\mathbf{y}}_1 = G_1(\mathbf{x}))$ to be close to the input $\mathbf{x}$. This can be formulated as a cycle-consistent loss:

$$\mathcal{L}_{cycle}(G_1, F_1, X_\downarrow) = \mathbb{E}_{\mathbf{x} \sim X\downarrow}\|\mathbf{x} - F_1(G_1(\mathbf{x}))\|_1. \tag{2}$$

Similarly, for a sample $\mathbf{y}_1 \in Y_\downarrow$, we use F_1 to perform translation, use D_{X_1} to calculate the adversarial loss, and then use G_1 to translate backward to calculate
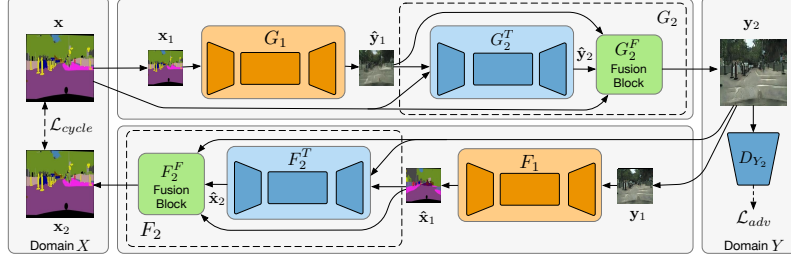
Fig. 3. Illustration of an overview of our Stage-2 for learning refining processes on the top of Stage-1's outputs. G_1 and F_1 are the translation networks learned in Stage-1. In the training process, we keep the weights of G_1 and F_1 fixed. Solid arrow denotes an input-output, and dashed arrow denotes a loss.

the cycle-consistent loss. Our full objective function for Stage-1 is a combination of the adversarial loss and the cycle-consistent loss:

$$\mathcal{L}_{Stage1} = \mathcal{L}_{adv}(G_1, D_{Y_1}, X_\downarrow, Y_\downarrow) + \mathcal{L}_{adv}(F_1, D_{X_1}, Y_\downarrow, X_\downarrow) \qquad (3)$$
$$+ \lambda[\mathcal{L}_{cycle}(G_1, F_1, X_\downarrow) + \mathcal{L}_{cycle}(F_1, G_1, Y_\downarrow)],$$

where λ denotes the weight of the cycle-consistent loss. We obtain the translations G_1 and F_1 by optimizing the following objective function:

$$G_1, F_1 = \arg\min_{G_1, F_1} \max_{D_{X_1}, D_{Y_1}} \mathcal{L}_{Stage1}, \qquad (4)$$

which encourages these translations to transform the results to another domain while preserving the intrinsic image content. As a result, the optimized translations G_1 and F_1 can perform a basic cross-domain translation in low resolution.

3.3   Stage-2: Refinement

Since it is difficult to learn a complicated translation with the limited ability of a single stage, the translated output of Stage-1 may seem plausible but still leaves us much room for improvement. To refine the output of Stage-1, we deploy Stage-2 with a stacked structure built on the top of the trained Stage-1 to complete the full translation to generate higher resolution results with finer details.

Stage-2 consists of two translation networks G_2, F_2 and two discriminator networks D_{X_2}, D_{Y_2}, as shown in Figure 3. We only describe the architecture of G_2, since F_2 shares the same design (see Figure 3).

G_2 consists of two parts: a newly initialized image translation network G_2^T and an adaptive fusion block G_2^F. Given the output of Stage-1 ($\hat{\mathbf{y}}_1 = G_1(\mathbf{x}_1)$), we use nearest up-sampling to resize it to match the original resolution. Different from the image translation network in Stage-1, which only takes $\mathbf{x} \in X$ as input, in Stage-2 we use both the current stage's input $\mathbf{x}$ and the previous stage's
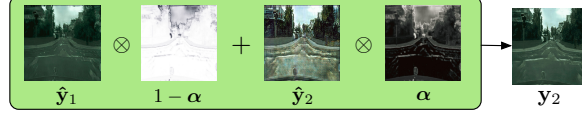
Fig. 4. Illustration of the linear combination in an adaptive fusion block. The fusion block applies the fusion weight map $\boldsymbol{\alpha}$ to find defects in the previous result $\hat{\mathbf{y}}_1$ and correct it precisely using $\hat{\mathbf{y}}_2$ to produce a refined output $\mathbf{y}_2$.

output $\hat{\mathbf{y}}_1$. Specifically, we concatenate $\hat{\mathbf{y}}_1$ and $\mathbf{x}$ along the channel dimension, and utilize G_2^T to obtain the refined result $\hat{\mathbf{y}}_2 = G_2^T(\hat{\mathbf{y}}_1, \mathbf{x})$.

Besides simply using $\hat{\mathbf{y}}_2$ as the final output, we introduce an adaptive fusion block G_2^F to learn a dynamic combination of $\hat{\mathbf{y}}_2$ and $\hat{\mathbf{y}}_1$ to fully utilize the entire two-stage structure. Specifically, the adaptive fusion block learns a pixel-wise linear combination of the previous results:

$$G_2^F(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2) = \hat{\mathbf{y}}_1 \odot (1 - \boldsymbol{\alpha}_x) + \hat{\mathbf{y}}_2 \odot \boldsymbol{\alpha}_x, \tag{5}$$

where $\odot$ denotes element-wise product and $\boldsymbol{\alpha} \in (0,1)^{H \times W}$ represents the fusion weight map, which is predicted by a convolutional network h_x:

$$\boldsymbol{\alpha}_x = h_x(\mathbf{x}, \hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2). \tag{6}$$

Figure 4 shows an example of adaptively combining the outputs from two stages.

Similar to Stage-1, we use a combination of adversarial and cycle-consistent losses to formulate our objective function of Stage-2:

$$\begin{aligned} \mathcal{L}_{Stage2} = {} & \mathcal{L}_{adv}(G_2 \circ G_1, D_{Y_2}, X, Y) + \mathcal{L}_{adv}(F_2 \circ F_1, D_{X_2}, Y, X) \\ & + \lambda \left[\mathcal{L}_{cycle}(G_2 \circ G_1, F_2 \circ F_1, X) + \mathcal{L}_{cycle}(F_2 \circ F_1, G_2 \circ G_1, Y) \right]. \end{aligned} \tag{7}$$

Optimizing this objective is similar to solving Equation 4. The translation networks G_2 and F_2 are learned to refine the previous results by correcting defects and adding details on them.

Finally, we complete our desired translations G and F by integrating the transformations in Stage-1 and Stage-2, which are capable of tackling a complex image-to-image translation problem under the unsupervised setting.

4 Experiments

The proposed approach is named *SCAN* or *SCAN Stage-N* if it has N stages in the following experiments. We explore several variants of our model to evaluate the effectiveness of our design in Section 4.7. In all experiments, we decompose the target translation into two stages, except for exploring the ability of the three-stage architecture in high-resolution tasks in Section 4.5.

We used the official released model of CycleGAN [34] and Pix2Pix [7] for 256×256 image translation comparisions. For 512×512 tasks, we train the CycleGAN with the official code since there is no available pre-trained model.

4.1   Network Architecture

For the image translation network, we follow the settings of [34,15], adopting the encoder-decoder architecture from Johnson *et al.* [8]. The network consists of two down-sample layers implemented by stride-2 convolution, six residual blocks and two up-sample layers implemented by sub-pixel convolution [20]. Note that different from [34], which used the fractionally strided convolution as the up-sample block, we use the sub-pixel convolution [20], for avoiding checkerboard artifacts [19]. The adaptive fusion block is a simple 3-layer convolutional network, which calculates the fusion weight map $\boldsymbol{\alpha}$ using two Convolution-InstanceNorm-ReLU blocks followed by a Convolution-Sigmoid block. For the discriminator, we use the PatchGAN structure introduced in [7].

4.2   Datasets

To demonstrate the capability of our proposed method for tackling the complex image-to-image translation problem under unsupervised settings, we first conduct experiments on the Cityscapes dataset [2]. We compare with the state-of-the-art approaches in the *Labels $\leftrightarrow$ Photo* task in 256×256 resolution. To further show the effectiveness of our method to learn complex translations, we also extended the input size to a challenging 512×512 resolution, namely the high-resolution Cityscapes *Labels $\rightarrow$ Photo* task.

Besides the *Labels $\leftrightarrow$ Photo* task, we also select six image-to-image translation tasks from [34], including *Map$\leftrightarrow$Aerial*, *Facades$\leftrightarrow$Labels* and *Horse$\leftrightarrow$Zebra*. We compare our method with the CycleGAN [34] in these tasks in 256×256 resolution.

4.3   Training Details

Networks in Stage-1 are trained from scratch, while networks in Stage-N are trained with the {Stage-1, $\cdots$, Stage-(N-1)} networks fixed. For the GAN loss, Different from the previous works [34,7], we adopt a gradient penalty term $\lambda_{gp}(||\nabla D(x)||_2-1)^2$ in the discriminator loss to achieve a more stable training process [12]. For all datasets, the Stage-1 networks are trained in 128×128 resolution, the Stage-2 networks are trained in 256×256 resolution. For the three-stage architecture in Section 4.5, the Stage-3 networks are trained in 512×512 resolution. We set batch size to 1, $\lambda = 10$ and $\lambda_{\mathrm{gp}} = 10$ in all experiments. All stages are trained with 100 epochs for all datasets. We use Adam [11] to optimize our networks with an initial learning rate as 0.0002, and decrease it linearly to zero in the last 50 epochs.

4.4   Evaluation Metrics

FCN Score and Segmentation Score. For the Cityscapes dataset, we adopt the FCN Score and the Segmentation Score as evaluation metrics from [7] for the *Labels $\rightarrow$ Photo* task and the *Photo $\rightarrow$ Labels* task, respectively. The FCN
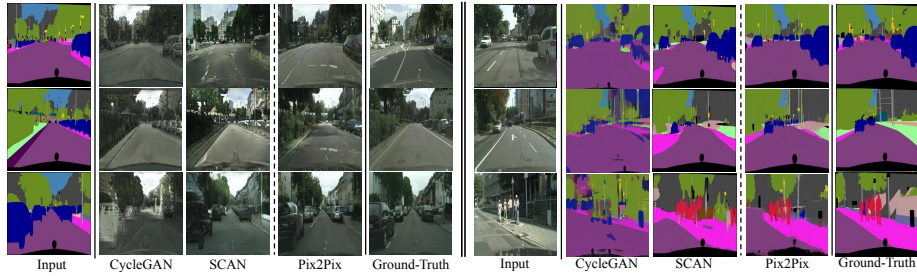
Input   CycleGAN   SCAN   Pix2Pix   Ground-Truth   Input   CycleGAN   SCAN   Pix2Pix   Ground-Truth

Fig. 5. Comparisons on the Cityscapes dataset of 256×256 resolution. The left sub-figure are *Labels $\rightarrow$ Photo* results and the right are *Photo $\rightarrow$ Labels* results. In the *Labels $\rightarrow$ Photo* task, our proposed SCAN generates more natural photographs than CycleGAN; in the *Photo $\rightarrow$ Labels* task, SCAN produces an accurate segmentation map while CycleGAN's results are blurry and suffer from deformation. SCAN also generates results that are visually closer to those of the supervised approach Pix2Pix than results of CycleGAN. Zoom in for better view.

Score employs an off-the-shelf FCN segmentation network [17] to estimate the realism of the translated images. The Segmentation Score includes three standard segmentation metrics, which are the per-pixel accuracy, the per-class accuracy, and the mean class accuracy, as defined in [17].

PSNR and SSIM. Besides using the FCN Score and the Segmentation Score, we also calculate the PSNR and the SSIM[25] for a quantitative evaluation. We apply the above metrics on the *Map $\leftrightarrow$ Aerial* task and the *Facades $\leftrightarrow$ Labels* task to measure both the color similarity and the structural similarity between the translated outputs and the ground truth images.

User Preference. We run user preference tests in the high-resolution Cityscapes *Labels $\rightarrow$ Photos* task and the *Horse$\rightarrow$Zebra* tasks for evaluating the realism of our generated photos. In the user preference test, each time a user is presented with a pair of results from our proposed SCAN and the CycleGAN [34], and asked which one is more realistic. Each pair of the results is translated from the same image. Images are all shown in randomized order. In total, 30 images from the Cityscapes test set and 10 images from the Horse2Zebra test set are used in the user preference tests. As a result, 20 participates make a total of 600 and 200 preference choices, respectively.

4.5   Comparisons

Cityscapes *Labels $\leftrightarrow$ Photo*. Table 1 shows the comparison of our proposed method SCAN and its variants with state-of-the-art methods in the Cityscapes *Labels $\leftrightarrow$ Photo* tasks. The same unsupervised settings are adopted by all methods except Pix2Pix, which is trained under a supervised setting.

On the FCN Scores, our proposed SCAN Stage-2 128-256 outperforms the state-of-the-art approaches considering the pixel accuracy, while being compet-

Table 1. FCN Scores in the Labels → Photo task and Segmentation Scores in the Photo → Labels task on the Cityscapes dataset. The proposed methods are named after *SCAN (Stage-1 resolution)-(Stage-2 resolution)*. *FT* means that we also *fine-tune* the Stage-1 model instead of fixing its weights. *FS* means directly training Stage-2 *from-scratch* without training the Stage-1 model.

Method	Labels → Photo			Photo → Labels		
	Pixel acc.	Class acc.	Class IoU	Pixel acc.	Class acc.	Class IoU
CycleGAN [34]	0.52	0.17	0.11	0.58	0.22	0.16
Contrast-GAN [15]	0.58	**0.21**	**0.16**	0.61	0.23	0.18
SCAN Stage-1 128	0.46	0.19	0.12	0.71	0.24	0.20
SCAN Stage-1 256	0.57	0.15	0.11	0.63	0.18	0.14
SCAN Stage-2 256-256	0.52	0.15	0.11	0.64	0.18	0.14
SCAN Stage-2 128-256 *FS*	0.59	0.15	0.10	0.36	0.10	0.05
SCAN Stage-2 128-256 *FT*	0.61	0.18	0.13	0.62	0.19	0.13
SCAN Stage-2 128-256	**0.64**	0.20	**0.16**	**0.72**	**0.25**	**0.20**
Pix2Pix [7]	0.71	0.25	0.18	0.85	0.40	0.32

itive considering the class accuracy and the class IoU. On the Segmentation Scores, SCAN Stage-2 128-256 outperforms state-of-the-art approaches in all metrics. Comparing SCAN Stage-1 256 with CycleGAN, our modified network yields improved results, which, however, still perform inferiorly to SCAN Stage-2 128-256. Also, we can find that SCAN Stage-2 128-256 achieves a much closer performance to the supervised approach Pix2Pix[7] than others.
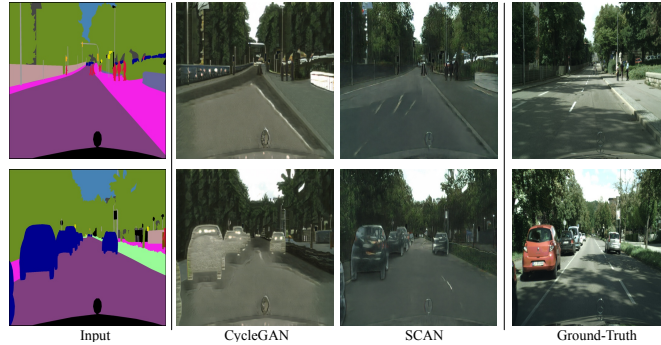
We also compare our SCAN Stage-2 128-256 with different variants of SCAN. Comparing SCAN Stage-2 128-256 with SCAN Stage-1 approaches, we can find a substantial improvement on the FCN Scores, which indicates that adding the Stage-2 refinement helps to improve the realism of the output images. On the Segmentation Score, comparison of the SCAN Stage-1 128 and SCAN Stage-1 256 shows that learning from low-resolution yields better performance. Comparison between the SCAN Stage-2 128-256 and SCAN Stage-1 128 shows that adding Stage-2 can further improve from the Stage-1 results. To experimentally prove that the performance gain does not come from merely adding model capacity, we conducted a SCAN Stage-2 256-256 experiments, which perform inferiorly to the SCAN Stage-2 128-256.

To further analyze various experimental settings, we also conducted our SCAN Stage-2 128-256 in two additional settings, including *leaning two stages from-scratch* and *fine-tuning Stage-1*. We add supervision signals to both stages for these two settings. Learning two stages from scratch shows poor performance in both tasks, which indicates joint training two stages together does not guarantee performance gain. The reason for this may lie in directly training a high-capacity generator is difficult. Also, fine-tuning Stage-1 does not resolve this problem and has smaller improvement compared with fixing weights of Stage-1.

To examine the effectiveness of the proposed fusion block, we compare it with several variants: 1) *Learned Pixel Weight* (LPW), which is our proposed fusion block; 2) *Uniform Weight* (UW), in which the two stages are fused with the same weight at different pixel locations $\hat{\mathbf{y}}_1(1-w) + \hat{\mathbf{y}}_2 w$, and during training w gradually increases from 0 to 1; 3) *Learned Uniform Weight* (LUW), which is

Table 2. FCN Scores and Segmentation Scores of several variants of the fusion block on the Cityscapes dataset.

Method	Labels → Photo			Photo → Labels		
	Pixel acc.	Class acc.	Class IoU	Pixel acc.	Class acc.	Class IoU
CycleGAN	0.52	0.17	0.11	0.58	0.22	0.16
SCAN 128-256 LPW	**0.64**	**0.20**	**0.16**	**0.72**	**0.25**	**0.20**
SCAN 128-256 UW	0.59	0.19	0.14	0.66	0.22	0.17
SCAN 128-256 LUW	0.59	0.18	0.12	0.70	0.24	0.19
SCAN 128-256 RF	0.60	0.19	0.13	0.68	0.23	0.18

Input   CycleGAN   SCAN   Ground-Truth

Fig. 6. Translation results in the *Labels → Photo* task on the Cityscapes dataset of 512×512 resolution. Our proposed SCAN produces realistic images that even look at a glance like the ground-truths. Zoom in for best view.

similar to *UW*, but w is a learnable parameter instead; 4) *Residual Fusion* (RF), which uses a simple residual fusion $\hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_2$. The results are illustrated in Table 2. It can be observed that our proposed LPW fusion yields the best performance among all alternatives, which indicates that the LPW approach can learn better fusion of the outputs from two stages than approaches with uniform weights.
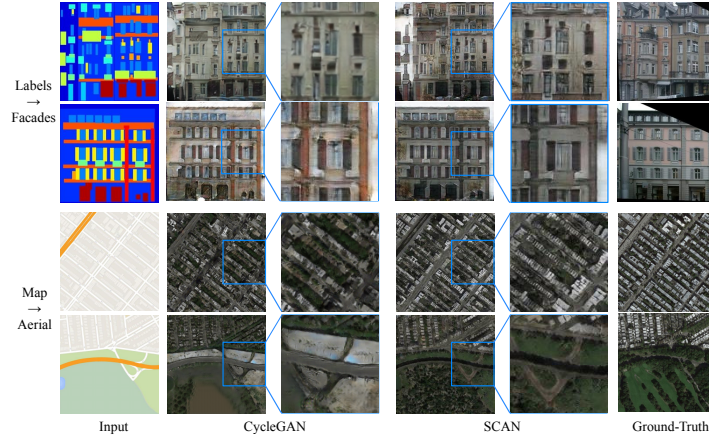
In Figure 5, we visually compare our results with those of the CycleGAN and the Pix2Pix. In the *Labels → Photo* task, SCAN generates more realistic and vivid photos compared to the CycleGAN. Also, the details in our results appear closer to those of the supervised approach Pix2Pix. In the *Photo → Labels* task, while SCAN can generate more accurate semantic layouts that are closer to the ground truth, the results of the CycleGAN suffer from distortion and blur.

High-Resolution Cityscapes *Labels → Photo*. The CycleGAN only considers images in 256×256 resolution, and results of training CycleGAN directly in 512×512 resolution are not satisfactory, as shown in Figure 1 and Figure 6.

By iteratively decomposing the Stage-2 into a Stage-2 and a Stage-3, we obtain a three-stage SCAN. During the translation process, the resolution of the output is growing from 128×128 to 256×256 and to 512×512, as shown in Figure 1. Figure 6 shows the comparison between our SCAN and the CycleGAN in the high-resolution Cityscapes *Labels → Photo* task. We can clearly see that

Table 3. PSNR and SSIM values in the *Map↔Aerial* and *Facades↔Labels* tasks.

Method	Aerial → Map		Map → Aerial		Facades → Labels		Labels → Facades	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CycleGAN[34]	21.59	0.50	12.67	0.06	6.68	0.08	7.61	0.11
SCAN	**25.15**	**0.67**	**14.93**	**0.23**	**8.28**	**0.29**	**10.67**	**0.17**

Input   CycleGAN   SCAN   Ground-Truth

Fig. 7. Translation results in the Labels→Facades task and the Aerial→Map task. Results of our proposed SCAN show finer details in both the tasks comparing with CycleGAN's results.

our proposed SCAN generates more realistic photos compared with the results of CycleGAN, and SCAN's outputs are visually closer to the ground truth images. The first row shows that our results contain realistic trees with plenty of details, while the CycleGAN only generates repeated patterns. For the second row, we can observe that the CycleGAN tends to simply ignore the cars by filling it with a plain grey color, while cars in our results have more details.

Also, we run a user preference study comparing SCAN with the CycleGAN with the setting described in Section 4.4. As a result, 74.9% of the queries prefer our SCAN's results, 10.9% prefer the CycleGAN's results, and 14.9% suggest that the two methods are equal. This result shows that our SCAN can generate overall more realistic translation results against the CycleGAN in the high-resolution translation task.

***Map↔Aerial* and *Facades↔Labels*.** Table 3 reports the performances regarding the PSNR/SSIM metrics. We can see that our methods outperform the CycleGAN in both metrics, which indicates that our translation results are more similar to ground truth in terms of colors and structures.

Figure 7 shows some of the sample results in the Aerial→Map task and the Labels→Facades task. We can observe that our results contain finer details while the CycleGAN results tend to be blurry.
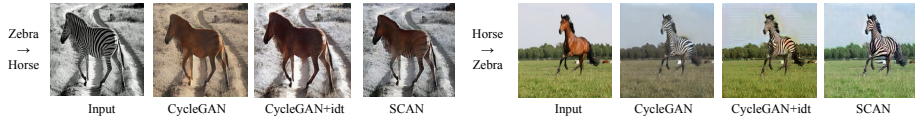
Fig. 8. Translation results in the Horse↔Zebra tasks. CycleGAN changes both desired objects and backgrounds. Adding an identity loss can fix this issue, but tends to be blurry compared with those from SCAN, which never uses the identity loss.
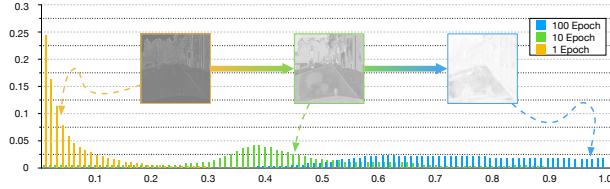

Fig. 9. Distributions of fusion weights over all pixels in different epochs. Each distribution is an average result over 1000 sample images from the Cityscapes dataset. Dashed arrows indicate average weights of fusion maps.

Horse↔Zebra. Figure 8 compares the results of SCAN against those of the CycleGAN in the Horse↔Zebra task. We can observe that both SCAN and the CycleGAN successfully translate the input images to the other domain. As the Figure 8 shows, the CycleGAN changes not only the desired objects in input images but also the backgrounds of the images. Adding the identity loss [34] can fix this problem, but the results still tend to be blurry compared with those from our proposed SCAN. A user preference study on Horse→Zebra translation is performed with the setting described in Section 4.4. As a result, 76.3% of the subjects prefer our SCAN's results against CycleGAN's, while 68.9% prefer SCAN's results against CycleGAN+idt's.

4.6   Visualization of Fusion Weight Distributions

To illustrate the role of the adaptive fusion block, we visualize the three average distributions of fusion weights ($\boldsymbol{\alpha}_x$ in Equation 5) over 1000 samples from Cityscapes dataset in epoch 1, 10, and 100, as shown in Figure 9. We observed that the distribution of the fusion weights gradually shifts from left to right. It indicates a consistent increase of the weight values in the fusion maps, which implies more and more details of the second stage are bought to the final output.

4.7   Ablation Study

In Section 4.5, we report the evaluation results of SCAN and its variants, here we further explore SCAN by removing modules from it:

Table 4. FCN Scores in the Cityscapes dataset for ablation study, evaluated on the *Labels → Photo* task with different variants of the proposed SCAN.

Method	Pixel acc.	Class acc.	Class IoU
SCAN Stage-1 128	0.457	0.188	0.124
SCAN Stage-2 128-256 w/o Skip,Fusion	0.513	0.186	0.125
SCAN Stage-2 128-256 w/o Skip	0.593	0.184	0.136
SCAN Stage-2 128-256 w/o Fusion	0.613	0.194	0.137
SCAN Stage-2 128-256	**0.637**	**0.201**	**0.157**

– SCAN w/o *Skip* Connection: remove the skip connection from the input to the translation network in the Stage-2 model , denoted by *SCAN w/o Skip*.
– SCAN w/o Adaptive *Fusion* Block: remove the final adaptive fusion block in the Stage-2 model , denoted by *SCAN w/o Fusion*.
– SCAN w/o *Skip* Connection and Adaptive *Fusion* Block: remove both the skip connection from the input to the translation network and the adaptive fusion block in the Stage-2 model , denoted by *SCAN w/o Skip, Fusion*.

Table 4 shows the results of the ablation study, in which we can observe that removing either the adaptive fusion block or the skip connection downgrades the performance. With both of the components removed, the stacked networks obtain marginal performance gain compared with Stage-1. Note that the fusion block only consists of three convolution layers, which have a relatively small size compared to the whole network. Refer to Table 1, in SCAN Stage-2 256-256 experiment, we double the network parameters compared to SCAN Stage-1 256, resulting in no improvement in the Label → Photo task. Thus, the improvement of the fusion block does not simply come from the added capacity.

Therefore, we can conclude that using our proposed SCAN structure, which consists of the skip connection and the adaptive fusion block, is critical for improving the overall translation performance.

5   Conclusions

In this paper, we proposed a novel approach to tackle the unsupervised image-to-image translation problem exploiting a stacked network structure with cycle-consistency, namely SCAN. The proposed SCAN decomposes a complex image translation process into a coarse translation step and multiple refining steps, and then applies the cycle-consistency to learn the target translation from unpaired image data. Extensive experiments on multiple datasets demonstrate that our proposed SCAN outperforms the existing methods in quantitative metrics and generates more visually pleasant translation results with finer details compared to the existing methods.

References

1. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: Proceedings of ICCV (2017)
2. 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: Proceedings of CVPR (2016)
3. Denton, E.L., Chintala, S., Fergus, R., et al.: Deep generative image models using a laplacian pyramid of adversarial networks. In: Proceedings of NIPS (2015)
4. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of ICCV (2015)
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of NIPS (2014)
6. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. ACM Transactions on Graphics (TOG) (2016)
7. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of CVPR (2017)
8. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of ECCV. pp. 694–711. Springer (2016)
9. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
10. Kim, T., Cha, M., Kim, H., Lee, J., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. Proceedings of ICML (2017)
11. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. Proceedings of ICLR (2014)
12. Kodali, N., Abernethy, J., Hays, J., Kira, Z.: On convergence and stability of gans. arXiv preprint arXiv:1705.07215 (2017)
13. Laffont, P.Y., Ren, Z., Tao, X., Qian, C., Hays, J.: Transient attributes for high-level understanding and editing of outdoor scenes. ACM Transactions on Graphics (TOG) (2014)
14. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. Proceedings of CVPR (2017)
15. Liang, X., Zhang, H., Xing, E.P.: Generative semantic manipulation with contrasting gan. Proceedings of NIPS (2017)
16. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. Proceedings of NIPS (2017)
17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of CVPR (2015)
18. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
19. Odena, A., Dumoulin, V., Olah, C.: Deconvolution and checkerboard artifacts. Distill **1**(10), e3 (2016)
20. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of CVPR (2016)
21. Simo-Serra, E., Iizuka, S., Sasaki, K., Ishikawa, H.: Learning to simplify: fully convolutional networks for rough sketch cleanup. ACM Transactions on Graphics (TOG) **35**(4), 121 (2016)

22. Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. Proceedings of ICLR (2016)
23. 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: Proceedings of CVPR (2018)
24. Wang, X., Gupta, A.: Generative image modeling using style and structure adversarial networks. In: Proceedings of ECCV (2016)
25. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing (TIP) **13**(4), 600–612 (2004)
26. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of ICCV (2015)
27. Yi, Z., Zhang, H., Gong, P.T., et al.: Dualgan: Unsupervised dual learning for image-to-image translation. Proceedings of ICCV (2017)
28. Xiong, W., Luo, W., Ma, L., Liu, W., Luo, J.: Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. Proceedings of CVPR (2018)
29. Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., Metaxas, D.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. Proceedings of ICCV (2016)
30. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Proceedings of ECCV (2016)
31. Zhang, R., Zhu, J.Y., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A.: Real-time user-guided image colorization with learned deep priors. ACM Transactions on Graphics (TOG) (2017)
32. Zhao, B., Chang, B., Jie, Z., Feng, J.: Modular generative adversarial networks. arXiv preprint arXiv:1804.03343 (2018)
33. Zhao, B., Wu, X., Cheng, Z.Q., Liu, H., Jie, Z., Feng, J.: Multi-view image generation from a single-view. arXiv preprint arXiv:1704.04886 (2017)
34. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of ICCV (2017)