applied sciences

MDPI

Article

Image-To-Image Translation Using a Cross-Domain Auto-Encoder and Decoder

Jaechang Yoo, Heesong Eom and Yong Suk Choi *

Department of Computer Science, Hanyang University, Seoul 04763, Korea; 3147@hanyang.ac.kr (J.Y.); heesong90@hanyang.ac.kr (H.E.)

* Correspondence: cys@hanyang.ac.kr

Received: 11 October 2019; Accepted: 5 November 2019; Published: 8 November 2019

Abstract: Recently, several studies have focused on image-to-image translation. However, the quality of the translation results is lacking in certain respects. We propose a new image-to-image translation method to minimize such shortcomings using an auto-encoder and an auto-decoder. This method includes pre-training two auto-encoders and decoder pairs for each source and target image domain, cross-connecting two pairs and adding a feature mapping layer. Our method is quite simple and straightforward to adopt but very effective in practice, and we experimentally demonstrated that our method can significantly enhance the quality of image-to-image translation. We used the well-known cityscapes, horse2zebra, cat2dog, maps, summer2winter, and night2day datasets. Our method shows qualitative and quantitative improvements over existing models.

Keywords: image-to-image translation; encoder-decoder; deep learning; feature mapping layer

1. Introduction

A variety of network models have been introduced to address the issue of image-to-image translation, which is a hot topic in the field of computer vision and graphics. Image-to-image translation involves learning the mapping from a source domain image to a target domain image. Early image-to-image translation methods used convolutional neural networks (CNN), which learn to minimize the loss of a pixel value between the source domain image and the target domain image but had the limitation of failing to produce more photorealistic images [1,2].

To solve this problem, a generative adversarial network (GAN) approach was recently suggested [3]. This approach used the information necessary to enhance the quality of the target translation image, rather than just using pixel-to-pixel dissimilarity, by additionally utilizing adversarial feedback to improve image translation quality [4–7].

Pix2pix [3] tried to solve the blurring problem, which is one of the problems faced by many CNN-based approaches, and had, thus, made various image translation tasks possible, such as changing a label (segmentation) image to an actual image, a daytime image to a night image, and a satellite map to a graphic map. Despite their promising results, there were still various problems, as depicted in Figure 1a. Furthermore, obtaining enough paired data to learn image translation was another difficult problem. For example, if we wanted to transform landscape photographs into paintings in the style of van Gogh, we might have to prepare real paintings by van Gogh for all the landscape pictures, which would be almost impossible in practice.

CycleGAN [8] enabled image-to-image translation even for unpaired image pairs and produced quite successful results. However, its unsupervised learning with unpaired data often causes instability of the images generated. When comparing the results of CycleGAN with the ground truth images of the cityscapes data, two different segments (building and tree) seemed to be confused with each other, as shown in Figure 1b. Such problems, as explained already, may raise the necessity for a more effective

translation method. In this paper, we propose a new translation method where the mapping process from source domain to target domain can work better, and thus suppress mistranslation by transferring the learning of domain and target feature spaces using both an auto-encoder and decoder. In contrast to commonly-used feature space learning methods [9–13], our approach utilizes an auto-encoder and an auto-decoder—which are pretrained for the source domain and target domain, respectively—and cross-connects them by mapping their feature spaces across the two domains.

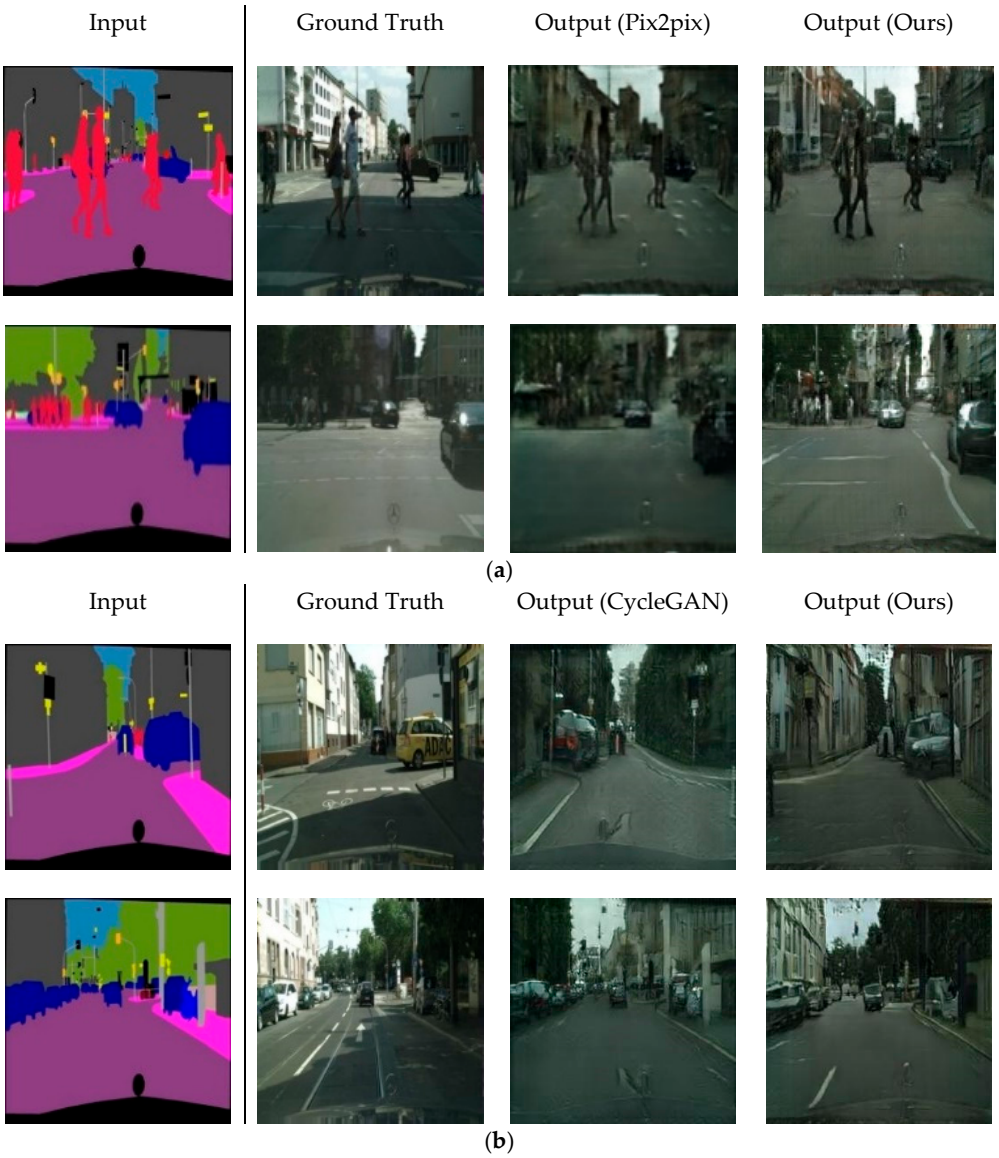

Figure 1. Low-quality photographs showing unstable flaws in the translated images of Pix2pix (**a**) and CycleGAN (**b**). From left to right: input, ground truth, output (baseline) and output (ours). (**a**) In the results of label → photo by Pix2pix, various notable noises and blurs are observed, which reduces the quality of the translated pictures. (**b**) In the translated pictures of label → photo by CycleGAN, some building segments are painted with tree pictures and vice versa, while our results show translating each segment accurately.

2. Related Work

2.1. Feature Extraction and Reconstruction

An auto-encoder is a method specialized in unsupervised feature learning. The auto-encoder consists of two parts: an encoder and a decoder. The encoder learns to translate the input into an

internal representation, which can also be used as a feature vector. The decoder learns to translate the internal representations translated by the encoder into the same form as the input.

Previous work has shown that extracting useful intermediate representations is effective for learning deep-generative or discriminative models. P. Vincent et al. [9,10] used corrupted inputs to train an auto-encoder for robust feature extraction. J. Masci et al. [11] introduced a hierarchical convolutional auto-encoder for a feature extraction. We use pre-trained auto-encoders to enable image translation between distinct feature spaces, such as "style" or "text," to transfer segmentation to photorealistic images. Details are described in Section 3.1.

2.2. Generative Adversarial Networks

Generative adversarial networks (GANs) [14] can learn to generate randomly sampled images with the idea of an adversarial loss that drives the output images of a generator indistinguishable from actual photographs. Recent methods adopted these networks to achieve magnificent results in several areas and applications [15–18]. Our work is also based on GANs for translating images but we focus to combine it with our training methodology.

2.3. GAN-Based Image-to-Image Translation

Recent studies have explored various models with GANs for image-to-image translation. For example, Pix2pix, the first GAN-based approach of P. Isola et al. [3], used a conditional, generative adversarial network [19] for image translation, when a source and target image pair was given. Similar methods have been adopted for several tasks, such as synthesizing a photograph from a sketch [20] and changing the weather of an original image [21]. More recently, T. park et al. [18] proposed a spatially-adaptive normalization layer to consider semantic information.

When paired training data were not available for some tasks, other studies have suggested several approaches to transform from an unpaired image. CoGAN [22] used a weight-sharing constraint to learn a joint representation over multi-domains. CycleGAN [8] used cycle consistency [23] and least-squares loss [6]. DualGAN [24] suggested a similar method to CycleGAN but used a Wasserstein loss [25] instead of the least-square loss. UNIT [26] and CD-GAN [27] used a shared latent space and cycle consistency for better quality and accuracy.

Unlike the approaches described above, our method is not limited to a specific task, nor do we rely on predefined relationships between the source and target domains. Our method can be applied to make a general-domain solution for many image-to-image translation tasks. We compare our trained model against previous approaches and experiment with our approach under various conditions in Section 4.

3. Method

3.1. Feature Space Mapping

We think that the decoder part can have the same significant effect as the encoder part and propose to use not only the pre-trained encoder but also a pre-trained decoder for the image-to-image translation task. We can consider an auto-encoder that has pre-trained feature extraction and reconstruction for the domain X. The encoder in this network, En_x, is responsible for representing the domain X to the latent space; En_x: $X \rightarrow Z_x$. Contrarily, the decoder De_x is responsible for representing the latent space to the domain X; De_x: $Z_x \rightarrow X$ (Figure 2). Our idea is using a "cross-connection" of two pairs of auto-encoders and decoders $En_x - De_x$ and $En_y - De_y$ for a generator of a GAN to improve image-to-image translation results. For example, we can connect En_x and De_y to make a generator to learn translation from domain X to domain Y. Since the encoder En_x has been pre-trained to represent the domain X to the latent space, it learns to extract the feature of domain X effectively in the image-to-image translation learning phase, and the decoder De_y also learns to reconstruct domain Y from the latent space while preserving common characteristics of Y with a pre-trained effect. This cross-connecting approach makes the results

of GAN more stable and photorealistic. Recently, like UNIT, image-to-image translation methods using auto encoders and decoders have been proposed. These methods try to train extracting latent space and translating domains in one training stage, resulting in some problems, such as low-quality translation, as we described in Section 4.3.2. In contrast, we think our method performs better because pre-training the auto-encoder and decoders provide more flexible advantages for the characteristics of each domains.

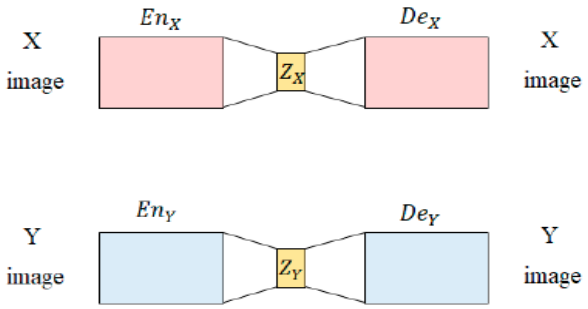

Figure 2. The architecture of an auto-encoder for pre-training the latent space of domain X and the latent space of domain Y.

To train auto-encoders for each domain, we follow the basic equation of the convolutional auto-encoder [11]. The representation of the k-th latent feature map extraction for input I is denoted as:

$$h^k = \sigma\,(I * W^k + b^k), \tag{1}$$

where sigma is an activation function, W_k is a weight, and bk is a bias, and * denotes the 2D convolution operation. The reconstruction of latent representation is obtained using

$$y = \sigma\,(\int_{k\in H} h^k * \hat{W} + c), \tag{2}$$

where c is a bias for input channel. H identifies the list of feature maps and $\hat{W}$ represents the 2D transposing convolution operation. Mean absolute error (MAE) or mean squared error (MSE) is used for a reconstruction loss between an output of the autoencoder and a ground-truth.

In addition to this, we construct an additional layer to combine two auto-encoder and decoder pairs. This layer we add learns to fill the "gap" between the space Z_x and Z_y. As we mentioned earlier, En_x learns to represent X relative to the latent space Z_x, and De_y reconstructs Y from space Z_y, so it does not guarantee that Z_x and Z_y, which are independently trained results of two different auto-encoders, lie in the same dimensional space. The added layer fills this gap to learn how to map two latent spaces properly.

After the cross-connection is complete (Figure 3), the entire network is fine-tuned by the learning method according to the model used to apply the cross-connected auto-encoder and decoder.

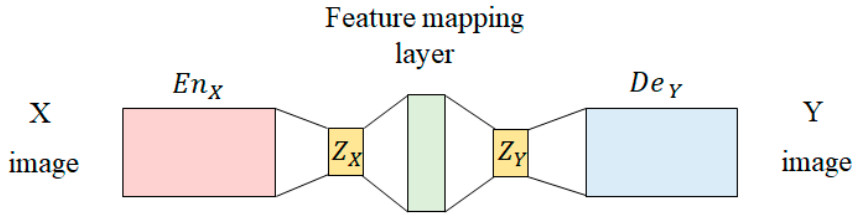

Figure 3. The concept of cross-connecting pre-trained autoencoders and decoders for feature space transfer (including a feature mapping layer).

3.2. Model Architecture

We refer to [28,29] for both generator and discriminator networks of CycleGAN [8] and Pix2pix [3]. These models are networks for following ResNet [30] or U-Net [31] architecture consisting of convolutional layers, normalization layers, skip-connections, and activation functions. A specific and detailed implementation is described below.

3.2.1. Generator and Discriminator

In a CycleGAN-based model, there are two translation processes: X → Y forward translation and Y → X reverse translation. To learn the unsupervised translation with this cycle consistency [23], two generator networks are needed: an En_x and De_y-connected pair for the forward translation process X → Y and an En_y and De_x pair for the reverse translation process Y → X. Each generator network is composed of processes in which the kernel-3, stride-2 convolutional layer, instance normalization layer, and ReLU-activation function. Each also contains processes of stride-2 deconvolutional layers after an encoded feature map passes through the multiple residual blocks. We use the PatchGAN [3] structure for the discriminator, which reduces the burden of calculation because it reduces the number of parameters in the network by classifying whether image patches are real or fake.

The Pix2pix-based model uses one generator consisting of En_x and De_y because there is a one-way translation process of X → Y. Further, PatchGAN is used for the discriminator, as in the CycleGAN-based model.

3.2.2. Feature Mapping Layer

Following Figure 4, the Pix2pix-based model uses a fully connected layer with one hidden layer as a mapping layer that maps the encoded latent vector to the target domain smoothly. For the CycleGAN-based generator in Figure 5, a convolutional layer with 3 × 3 size kernel and stride-1 was used for the feature mapping layer. Due to the structural limitations of the residual block used in CycleGAN, we use the convolutional layer for the large size feature map as a substitute for the fully connected layer, which is used as the feature mapping layer in the Pix2pix-based model. For an effective image translation, skip-connection in the Pix2pix (or residual connection in the CycleGAN) is removed from the feature mapping layer, while the connection is preserved in other layers, as shown in Figures 4 and 5.

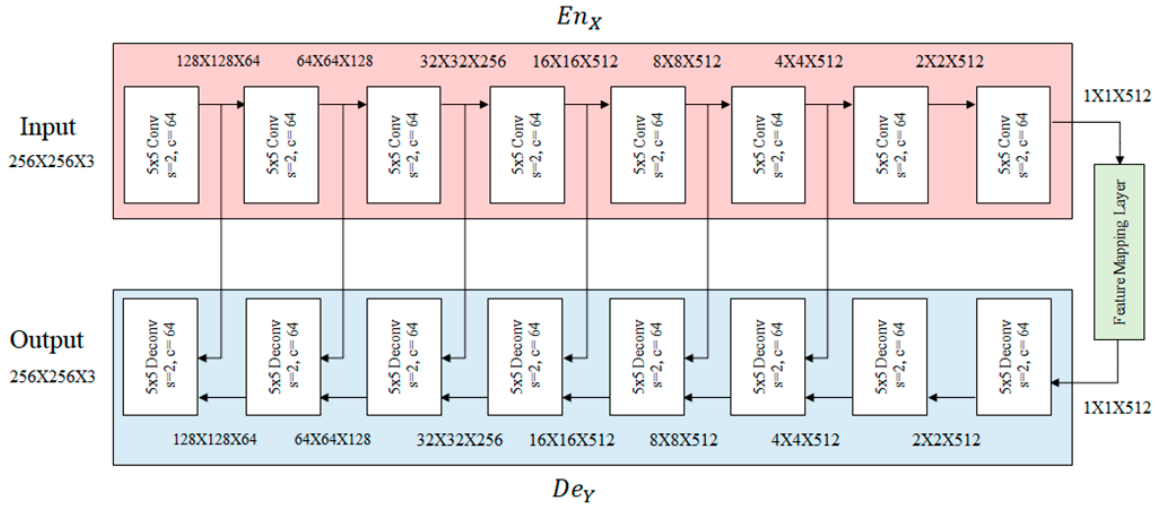

Figure 4. The generator architecture of our model based on Pix2pix. Cross-connected structure of the pre-trained encoder and decoder with feature mapping layer and skip-connection of U-net.

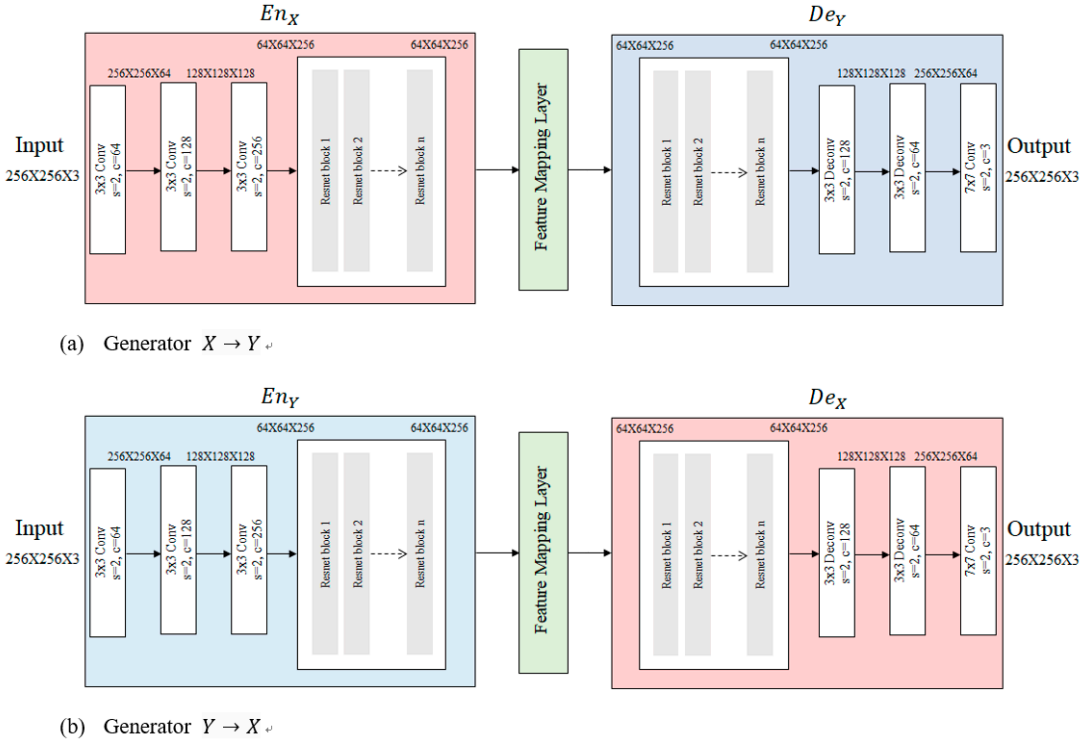

Figure 5. CycleGAN-based generator architecture of our model. Cross-connected structure of the pre-trained encoder and decoder with feature mapping layer and residual connection of residual block. (**a**) Generator in our model for translating images from domain X to Y. (**b**) Generator in our model for translating images from domain Y to X.

4. Results

We compared our method to others under various experimental conditions for both unpaired and paired image-to-image translations. For comparison, original models of Pix2pix and CycleGAN were used for baselines to evaluate the performance of our method for each paired and unpaired image-to-image translation task. UNIT was also trained and used for the comparison with our method.

4.1. Datasets and Evaluation Metrics

First, we used the cityscapes dataset [32] presented in a previous work for a quantitative comparison. The cityscapes dataset contains various distributions of pairs for semantic segmented images and photorealistic images of city pictures taken on the road, and it is a common and challenging dataset used to test image-to-image translation performance [3,8]. We used 2975 image pairs for training and 500 pairs for testing. We also tested our approach on the horse2zebra dataset to see how it affects the qualitative performance of object transfiguration. We used 1067 horse images and 1334 zebra images for training, and 120 horse images and 140 zebra images for testing. Cat2dog, maps (satellite map), summer2winter, and night2day datasets were also used for qualitative comparisons.

For measuring metrics to evaluate our models, we referred to the metrics used in existing methods: mean absolute error (MAE), mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), inception score (IS), and Fréchet inception distance (FID score). MAE and MSE are the most common ways of measuring the accuracy of the result images in computer vision tasks [15,33]. In some cases, however, these are not appropriate for evaluating the results of the GAN approach, and thus, various methods use PSNR or SSIM to assess the quality of the GAN results quantitatively [17,34–37]. Recent methods also used the inception score [16,38], or FID score [16,39] to measure the visual quality and diversity of generated images. In this work, we used all six metrics for evaluation.

4.2. Implementation Details

For the paired translation task, we used the U-Net based generator model that is used in the Pix2pix model. For the unpaired translation task, we adopted the ResNet-based architecture used in CycleGAN. Both models use a 70 × 70 PatchGAN for discriminators. An objective function combined with adversarial loss and cycle consistency loss is used for the CycleGAN-based model. Similarly, the Pix2pix-based model uses an objective function that combines an adversarial loss with a reconstruction loss. The coefficients of loss values were set as lcyc = 10 and lpix2pix = 100. According to the guidelines of baselines, the Adam optimizer was used and the learning rate parameters of both models were set as 0.0002. In the experiment with the cityscapes dataset, approximately 100 epochs of training were performed, and in the experiment with the horsse2zebra dataset, 200 epochs were performed, all in the same way as the learning conditions for Pix2pix and CycleGAN.

4.3. Qualitative Comparisons

4.3.1. Comparisons with Baselines

We applied our method on CycleGAN and evaluated our results qualitatively. Figures 6 and 7 demonstrate that our model solves the critical problems found in the mistranslated images generated by the baseline model (CycleGAN). Figure 6 shows that some segmented objects are translated with incorrect textures by the baseline model, but our model translates the segmented images correctly on the cityscapes task. Figure 7 shows that some regions or objects other than horses are painted partly with zebra patterns or poorly translated. In the case of our model, it produces better quality images with the correct translation on the horse2zebra task.

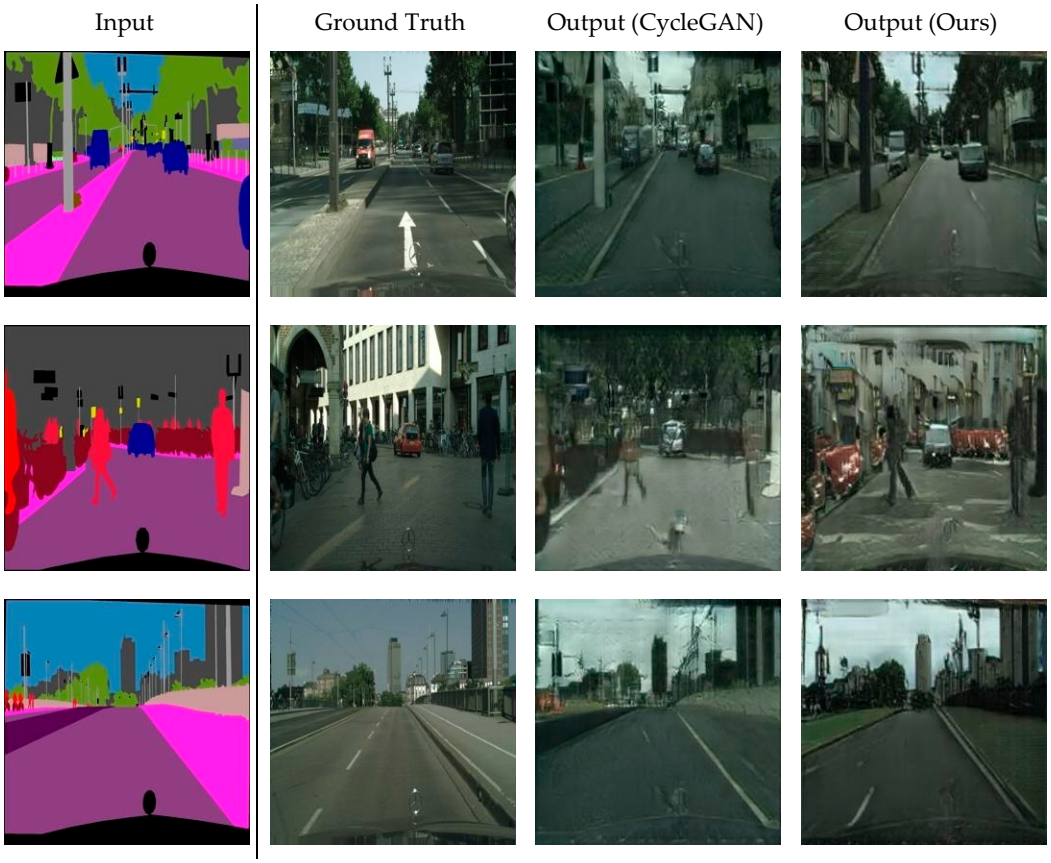

Figure 6. *Cont*.

Figure 6. Translated pictures of CycleGAN and our method for the cityscapes dataset. From left to right: input, ground truth, Cycle-GAN (Baseline), and our method. The results of CycleGAN show unstable translations, including mistranslations between colonnade tree label and the building appearance. Our method helps to avoid these instabilities during the learning stage and produces more correct translation results.

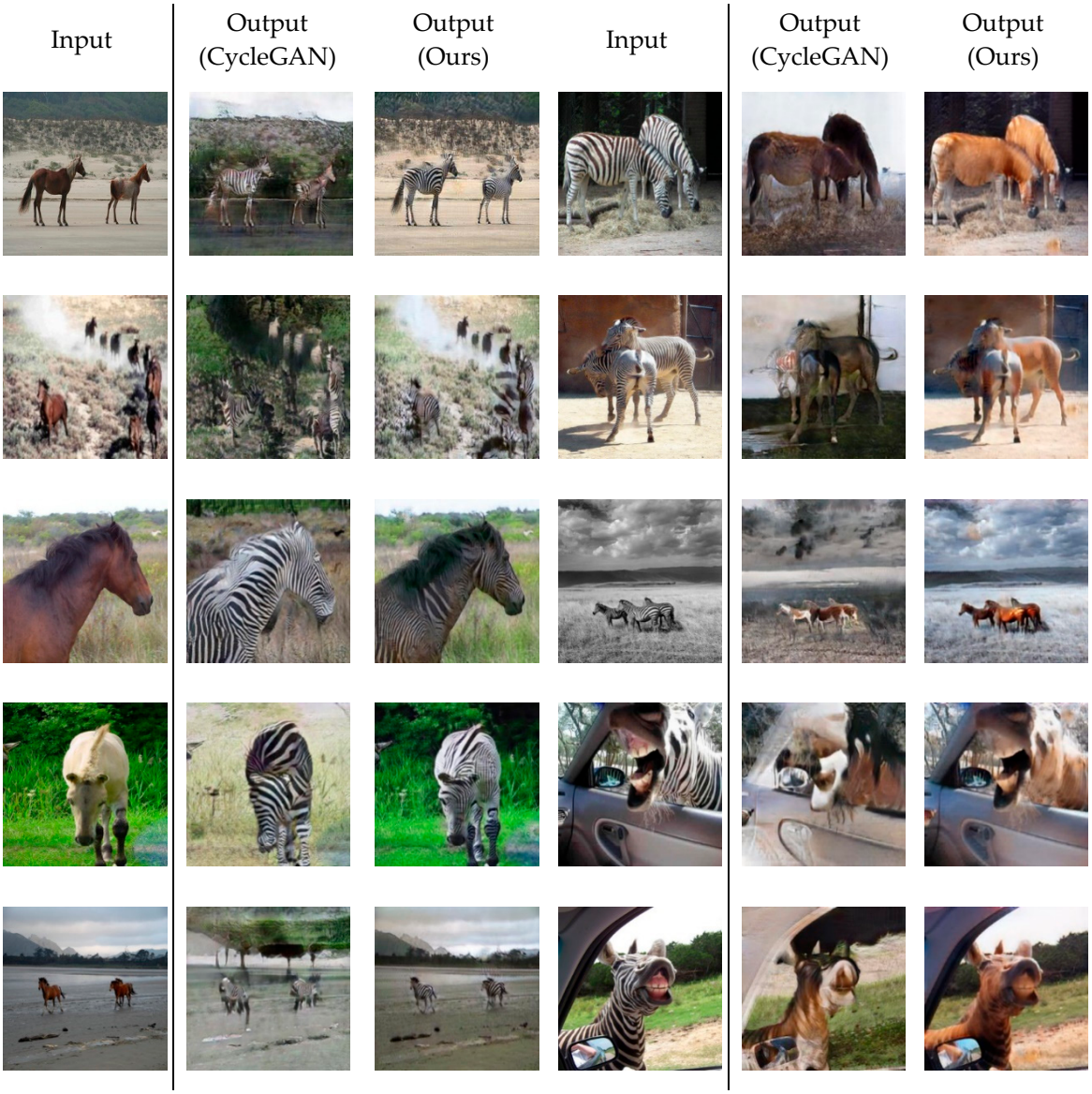

Figure 7. Comparison of pictures translated by CycleGAN and our method from the horse2zebra dataset. CycleGAN results show mistranslations such as changing texture of background, or removing detailed characteristics of the source object, or just inverting image colors. The results of our method show such mistranslations can instead be correctly translated.

Figure 8 shows more details of the resultant translation images. It shows that our model can resolve the instabilities of the baseline model (CycleGAN) effectively. Particularly, our model translates the horses to zebras only, and preserves the original features of the other parts of the image (background and other objects) that should not be translated. Since the horse2zebra dataset is unpaired between the horse domain and the zebra domain, quantitative comparisons with ground-truth are not possible, and qualitative comparisons are the only indicator of the quality of the results.

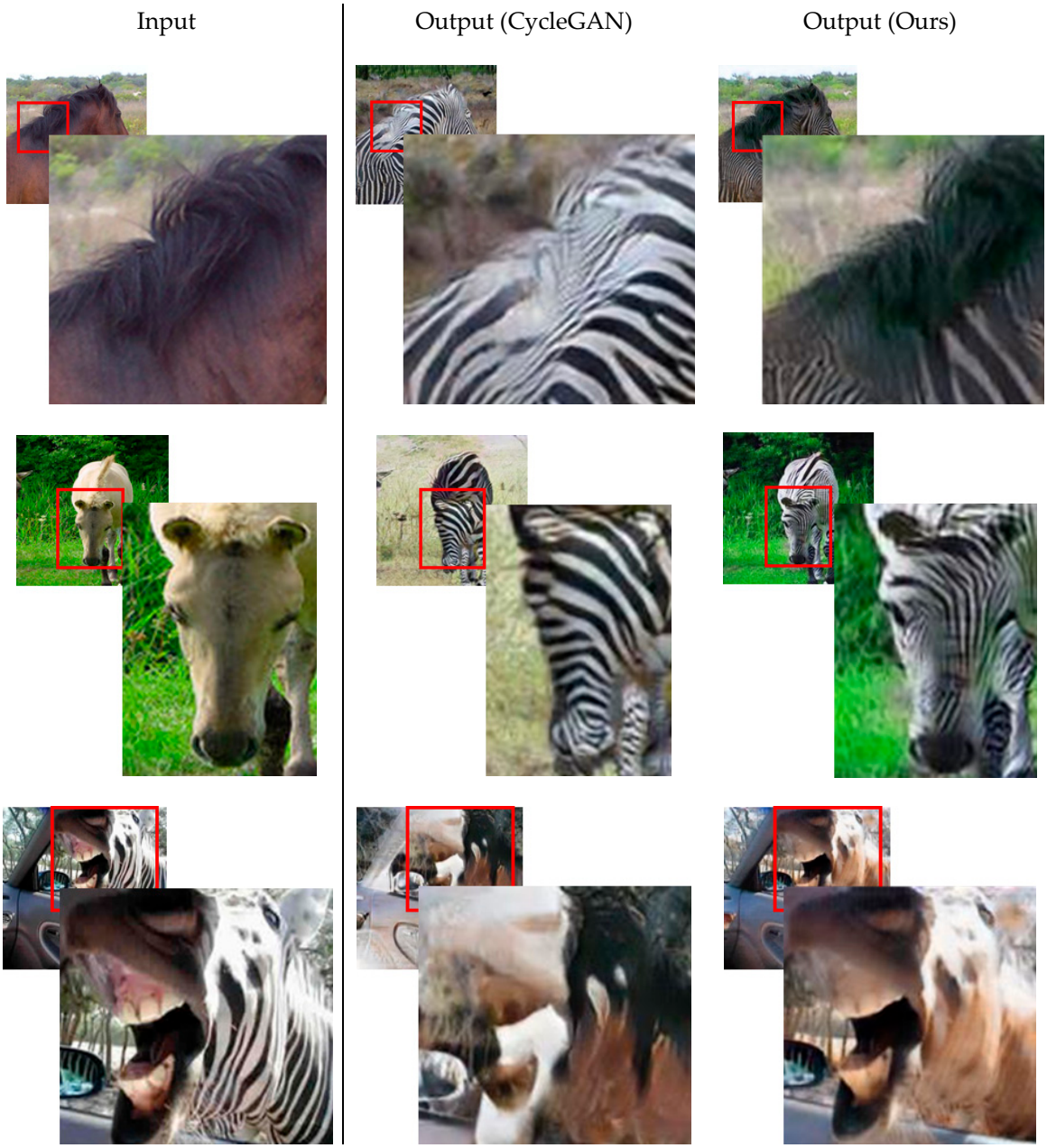

Figure 8. Detailed portions of translated pictures. Enlargements of the results of the baseline (CycleGAN) reveal mistranslation from the horse hair to zebra patterns, elimination of detailed features like eyes and nose, and inverting (or blurring) colors. Our model is more stable than the baseline because it makes translations while preserving the main characteristics of all the objects in the image, including eyes, nose, hair, and colors.

We also applied our method on Pix2pix and evaluated our results qualitatively. In Figure A1, Pix2pix shows the mixed textures, faded edges, and blurred objects of translated images on the

cityscapes dataset, but they are translated with, apparently, relatively sharp edges, and distinct textures and objects by our method.

4.3.2. Comparisons with UNIT

Figures 9 and A2 show that our model (applied on CycleGAN) generates better quality images than UNIT on different datasets. UNIT effectively translates complex features to simple features, including distinguishing boundaries and classifying objects. When generating complex features, however, the results of UNIT show some limitations, such as cracked or over-sharpened textures. On the summer2winter and night2day datasets, UNIT shows the surface of an object that appears to be cracked and the existence of an object is lost in severe cases. Our model, on the other hand, translates only the parts of the image that need to be changed to map the characteristics of these datasets. In summer2winter datasets, our model is good at finding areas where snow accumulates in winter images and translates them into images suitable for the summer season.

On the cat2dog dataset, UNIT tries to translate the image while maintaining the overall shape of input cat or dog images because of their shared latent space. We think that learning to force the characteristics of different domains into the same latent space is not an efficient method for domain translation, as we can see in the results. In the case of map datasets, our model not only shows the proper translation of labels, but also the detailed appearance of a building, whereas the UNIT model fails to restore objects such as buildings or trees in the area and mistranslates a label. Most results of unpaired and paired datasets show that our model's performance is better than the UNIT model because our model learns to map the feature space between source and target domains, allowing for more versatile and realistic domain translation.

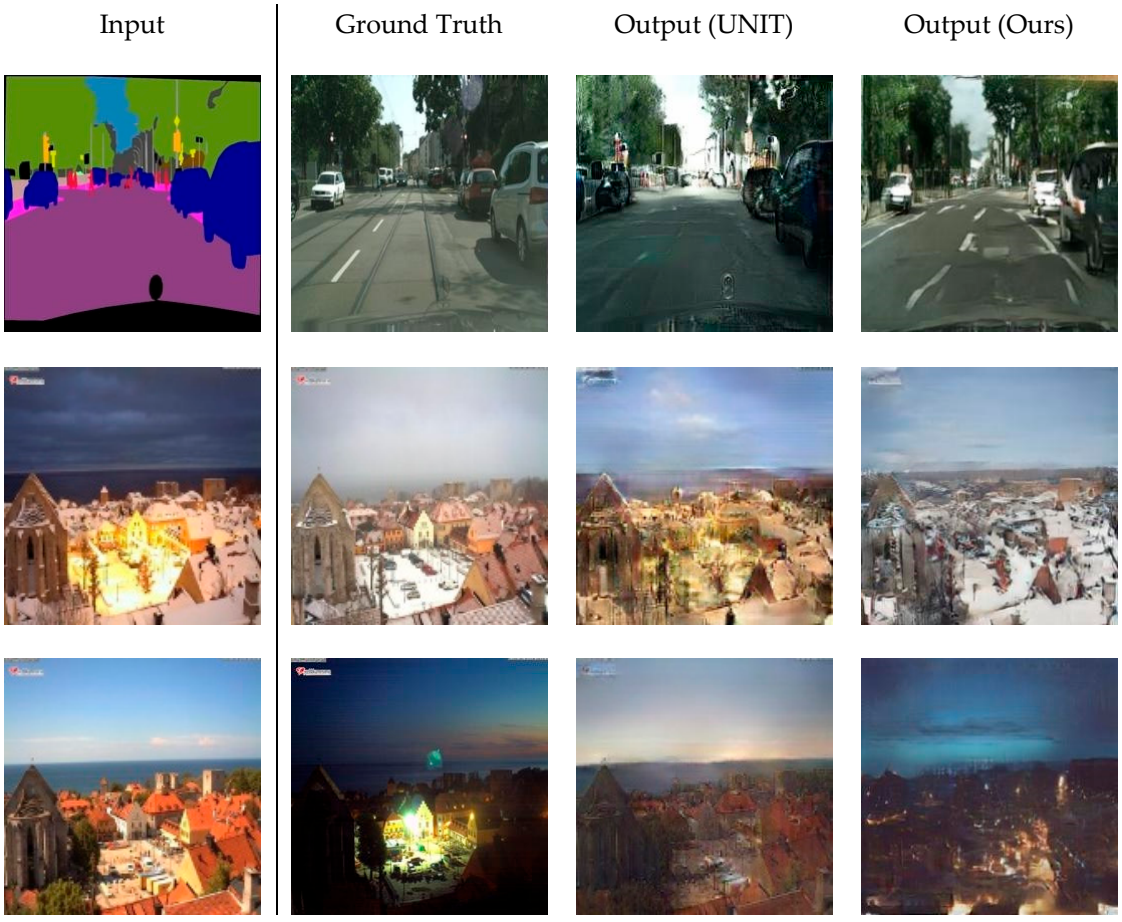

Figure 9. *Cont.*

Figure 9. Comparison of the UNIT model and our model on paired datasets. On the cityscapes and night2day datasets, UNIT generates images with overly sharp textures, while our model generates a smoother, more natural texture. Additionally, UNIT results on the maps dataset show a missing "park" (green class in the input) section and a lower quality of building details than our method.

4.4. Quantitative Comparisons

We also report comparisons of quantitative evaluations here. Table 1 shows that our model makes improvements over other baseline models, showing a lower FID and a higher SSIM and inception score. In particular, the inception score of our model is improved by about 6% and the FID of our model is improved by about 8% in the label → photo task compared to the baseline model (CycleGAN). This result is particularly notable because, as we mentioned earlier, previous metrics (such as MAE, MSE, and SSIM) had some issues in measuring the GAN results of photorealistic image generation. We focus more on IS and FID scores (especially on FID) in this task because PSNR is also a measurement obtained from using MAE and MSE.

Table 1. Evaluation of baseline models and our model on the cityscapes dataset (label → photo). Higher value is better in SSIM, PSNR, IS. Lower value is better in MAE, MSE, FID.

Metrics	**Label → Photo**					
	SSIM ↑	**PSNR ↑**	**MAE ↓**	**MSE ↓**	**IS ↑**	**FID ↓**
Baseline (Pix2pix)	0.2863	12.8684	41.7085	3547.5539	2.3955 ± 0.1610	101.04794
Baseline (CycleGAN)	0.3416	14.5333	35.4496	**2558.5918**	2.3421 ± 0.2216	75.3792
UNIT	0.3526	**14.6102**	**35.4176**	2653.8875	2.3304 ± 0.2486	82.7469
Ours + 3*3 convolutional layer (on CycleGAN)	**0.3571**	14.2083	36.1383	2714.3276	**2.4554 ± 0.1582**	**69.1041**

4.5. The Effectiveness of the Feature Mapping Layer

Table 2 shows the effectiveness of adding the feature mapping layer between the auto-encoder of the source domain and the auto-decoder of the target domain on the photo → label translation of the cityscapes dataset. We experiment using different feature mapping layers with various settings: without feature mapping layer, 1×1 convolutional layer, 3×3 convolutional layer, 7×7 depth-wise convolutional layer, 1×1 convolutional layer with skip-connection, and 3×3 convolutional layer with residual connection for our CycleGAN-based model.

Table 2. Ablation study of our model on cityscapes dataset (photo → label). Higher value is better in SSIM, PSNR, IS. Lower value is better in MAE, MSE, FID.

Metrics	Photo → Label					
	SSIM ↑	PSNR ↑	MAE ↓	MSE ↓	IS ↑	FID ↓
Baseline (CycleGAN)	**0.5568**	15.0977	**25.7486**	2185.6808	2.1170 ± 0.1927	107.4895
Ours w/o feature mapping layer	0.5054	13.5550	37.9918	2999.5845	**2.2096 ± 0.1290**	107.3393
Ours + 1×1 convolutional layer	0.5326	14.5956	33.2103	2439.4565	1.9608 ± 0.1521	96.7853
Ours + 3×3 convolutional layer	0.5526	**15.4477**	29.9302	**2040.3728**	2.0509 ± 0.1997	**93.4251**
Ours + 7×7 depth-wise convolutional layer	0.5029	13.1269	37.9808	3301.7309	2.1264 ± 0.1557	126.5632
Ours + 1×1 convolutional layer + residual connection (in FML)	0.5006	13.6503	39.0486	2928.5135	2.1674 ± 0.1042	120.7020
Ours + 3×3 convolutional layer + residual connection (in FML)	0.5017	13.5443	37.6828	3026.7782	2.0445 ± 0.1418	108.9408

Our model with a 3×3 convolutional layer shows high overall performance, especially in terms of FID score, which is considered an important indicator. The 1×1 convolutional layer also shows better overall performance than most other experiments in FID. Using only a pre-trained auto-encoder and decoder shows the best performance in IS but shows poor results with respect to all other metrics. These results indicate that our feature mapping layer helps with learning to map latent space when used with the suitable pre-trained encoder and decoder, which results in improving the translation quality significantly.

Adding skip-connection in the feature mapping layer shows even worse results compared to the baseline model. We think that connecting encoder and decoder of two different domains interferes with learning correct translation.

5. Discussion

We introduced an image-to-image translation method that uses a pre-trained auto-encoder and decoder to enhance the quality of translation. First, we trained two distinct auto-encoders for the source domain X and target domain Y, respectively. Then, we cross-connected the pre-trained auto-encoder (for the feature extractor of X) and auto-decoder (for the reconstructor of Y), adding a feature mapping layer. In experiments, we showed that our method improves the quality of translation results significantly in comparison with the existing baselines. We also investigated how to add a feature mapping layer in order to further enhance the performance of our model. Despite qualitative and quantitative improvements identified in experiments, our method is limited to unimodal image-to-image translation. We plan to investigate more sophisticated feature mapping architectures between the auto-encoder and decoder by conducting experiments on multimodal tasks.

Author Contributions: Conceptualization and methodology, J.Y., H.E., and Y.S.C.; investigation, J.Y. and H.E.; data curation and conceiving experiments, J.Y.; performing experiments and designing results, H.E.; writing—original draft preparation, J.Y. and H.E.; writing—review and editing, J.Y., H.E., and Y.S.C.

Funding: This work was supported by the National Research Foundation of Korea (NRF), a grant funded by the Korean government (Ministry of Science and ICT) (2018R1A5A7059549), the ITRC (Information Technology Research Center) support program (IITP-2017-0-01642) supervised by the IITP (Institute for Information and communications Technology Promotion, Korea), and the Technology Innovation Program (10077553) funded by the Ministry of Trade, Industry, and Energy (MOTIE, Korea).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

We provide our source code at https://github.com/Rooooss/I2I_CD_AE.

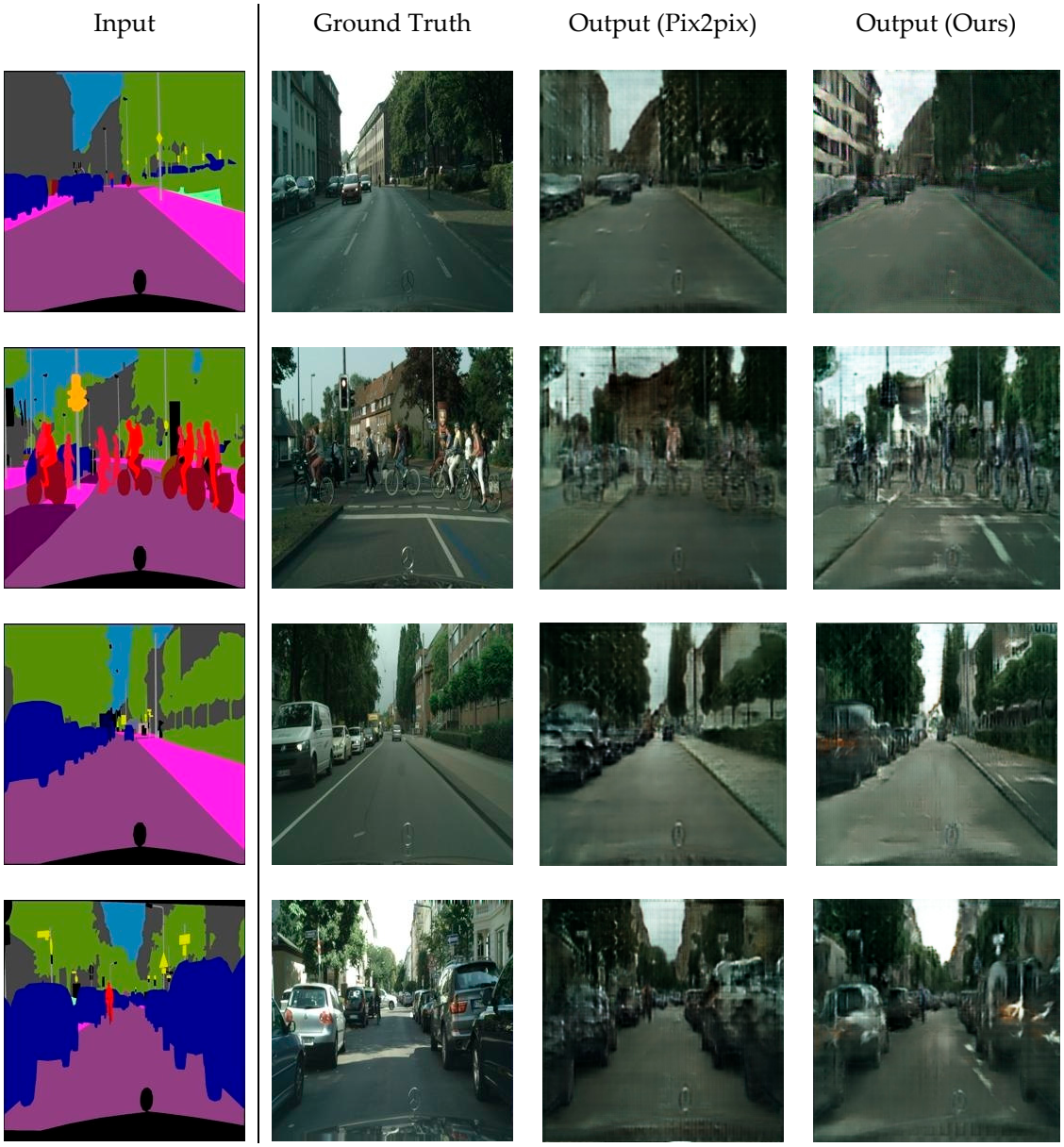

Figure A1. Comparison of Pix2pix and our model. The introduction of noise, including mixed textures, faded edges, and blurred objects, are mitigated in our model.

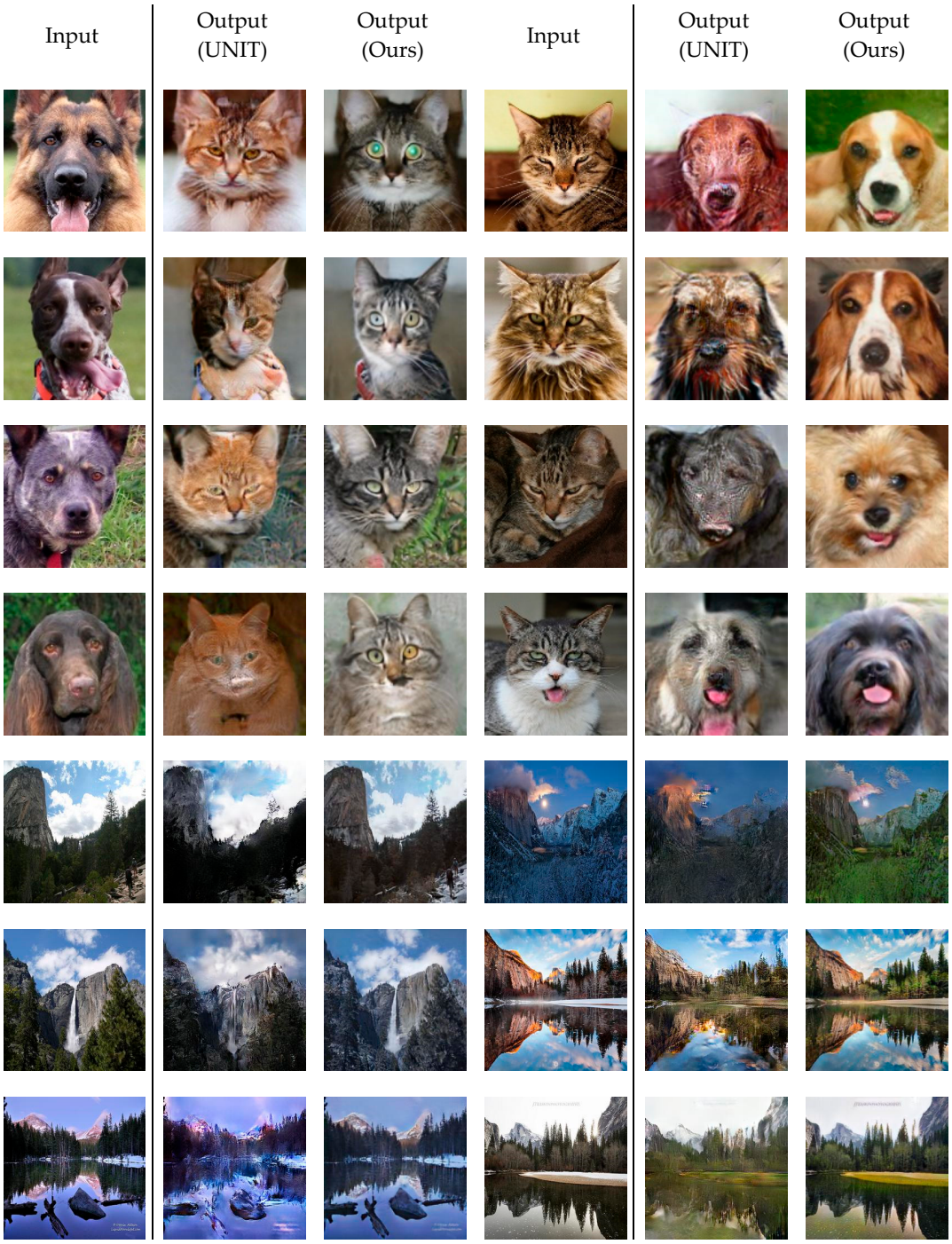

Figure A2. Comparison of the UNIT model and our model on unpaired datasets. Our results show more promising translation results on the cat2dog and summer2winter datasets compared to the baseline model (UNIT).

References

1. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context Encoders: Feature Learning by Inpainting. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
2. Zhang, R.; Isola, P.; Efros, A.A. Colorful Image Colorization. In Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 649–666.

3. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-To-Image Translation with Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976.
4. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. In Proceedings of the 34th International Conference on Machine Learning-Volume, Sydney, Australia, 6–11 August 2017; pp. 1857–1865.
5. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. Stackgan: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5907–5915.
6. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul Smolley, S. Least Squares Generative Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2794–2802.
7. Zhu, J.Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A.A.; Wang, O.; Shechtman, E. Toward Multimodal Image-to-Image Translation. In Proceedings of the 32th Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 465–476.
8. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
9. Vincent, P.; Larochelle, H.; Bengio, Y.; Manzagol, P.A. Extracting and Composing Robust Features with Denoising Autoencoders. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 1096–1103.
10. Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.A. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371–3408.
11. Masci, J.; Meier, U.; Cire San, D.; Schmidhuber, J. Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. In Proceedings of the 21st International Conference on Artificial Neural Networks, Espoo, Finland, 14–17 June 2011; pp. 52–59.
12. Rifai, S.; Vincent, P.; Muller, X.; Glorot, X.; Bengio, Y. Contractive Auto-Encoders: Explicit Invariance during Feature Extraction. In Proceedings of the 28th International Conference on Machine Learning, Washington, DC, USA, 28 June–2 July 2011; pp. 833–840.
13. Kingma, D.P.; Welling, M. Auto-Encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
14. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the 29th Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
15. Nie, D.; Trullo, R.; Lian, J.; Petitjean, C.; Ruan, S.; Wang, Q.; Shen, D. Medical Image Synthesis with Context-Aware Generative Adversarial Networks. In Proceedings of the 20th International Conference on Medical Image Computing and Computer-Assisted Intervention, Quebec City, QC, Canada, 11–13 September 2017; pp. 417–425.
16. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.
17. Kupyn, O.; Budzan, V.; Mykhailych, M.; Mishkin, D.; Matas, J. DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8183–8192.
18. Park, T.; Liu, M.Y.; Wang, T.C.; Zhu, J.Y. Semantic image synthesis with spatially-adaptive normalization. *arXiv* **2019**, arXiv:1903.07291.
19. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
20. Sangkloy, P.; Lu, J.; Fang, C.; Yu, F.; H, J. Scribbler: Controlling Deep Image Synthesis with Sketch and Color. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5400–5409.
21. Zhang, H.; Sindagi, V.; Patel, V.M. Image de-raining using a conditional generative adversarial network. *arXiv* **2017**, arXiv:1701.05957. [CrossRef]
22. Liu, M.Y.; Tuzel, O. Coupled Generative Adversarial Networks. In Proceedings of the 31th Advances In Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 469–477.

23. Zhou, T.; Krahenbuhl, P.; Aubry, M.; Huang, Q.; Efros, A.A. Learning Dense Correspondence via 3d-Guided Cycle Consistency. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 117–126.
24. Yi, Z.; Zhang, H.; Tan, P.; Gong, M. Dualgan: Unsupervised Dual Learning for Image-to-Image Translation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2849–2857.
25. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein gan. *arXiv* **2017**, arXiv:1701.07875.
26. Liu, M.Y.; Breuel, T.; Kautz, J. Unsupervised Image-to-Image Translation Networks. In Proceedings of the 32th Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 700–708.
27. Yang, X.; Xie, D.; Wang, X. Crossing-Domain Generative Adversarial Networks for Unsupervised Multi-Domain Image-to-Image Translation. In Proceedings of the 26th ACM Multimedia Conference on Multimedia Conference, Seoul, Korea, 22–26 October 2018; pp. 374–382.
28. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
29. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 694–711.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
31. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
32. 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 the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3213–3223.
33. Silva, E.A.; Panetta, K.; Agaian, S.S. Quantifying image similarity using measure of enhancement by entropy. In Proceedings of the Mobile Multimedia/Image Processing for Military and Security Applications, Orlando, FL, USA, 11–12 April 2007; p. 65790U.
34. Kim, J.; Lee, J.K.; Lee, K.M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1646–1654.
35. Kim, J.; Lee, J.K.; Lee, K.M. Deeply-Recursive Convolutional Network for Image Super-Resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1637–1645.
36. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 624–632.
37. Galteri, L.; Seidenari, L.; Bertini, M.; Del Bimbo, A. Deep Generative Adversarial Compression Artifact Removal. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4826–4835.
38. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved Techniques for Training Gans. In Proceedings of the 31th Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 2234–2242.
39. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochriter, S. Gans Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Proceedings of the 32th Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6626–6637.