electronics

MDPI

Article

Accurate and Consistent Image-to-Image Conditional Adversarial Network

Naeem Ul Islam [1], Sungmin Lee [1,2] and Jaebyung Park [1,2,*]

[1] Core Research Institute of Intelligent Robots, Jeonbuk National University, 567 Baekje-daero, Deokjin-gu, Jeonju-si, Jeollabuk-do 54896, Korea; naeem@jbnu.ac.kr (N.U.I.); leesungmin@jbnu.ac.kr (S.L.)

[2] Division of Electronics and Information Engineering, Jeonbuk National University, 567 Baekje-daero, Deokjin-gu, Jeonju-si, Jeollabuk-do 54896, Korea

* Correspondence: jbpark@jbnu.ac.kr; Tel.: +82-63-270-4283

Received: 29 January 2020; Accepted: 25 February 2020; Published: 27 February 2020

Abstract: Image-to-image translation based on deep learning has attracted interest in the robotics and vision community because of its potential impact on terrain analysis and image representation, interpretation, modification, and enhancement. Currently, the most successful approach for generating a translated image is a conditional generative adversarial network (cGAN) for training an autoencoder with skip connections. Despite its impressive performance, it has low accuracy and a lack of consistency; further, its training is imbalanced. This paper proposes a balanced training strategy for image-to-image translation, resulting in an accurate and consistent network. The proposed approach uses two generators and a single discriminator. The generators translate images from one domain to another. The discriminator takes the input of three different configurations and guides both the generators to generate realistic images in their corresponding domains while ensuring high accuracy and consistency. Experiments are conducted on different datasets. In particular, the proposed approach outperforms the cGAN in realistic image translation in terms of accuracy and consistency in training.

Keywords: generative adversarial network; convolutional neural network; consistent image-to-image translation network; autoencoders

1. Introduction

Translating images between different domains has many important applications in the field of robotics and computer vision, including terrain shape estimation, tip-over and collision avoidance, scene understanding, image colorization, styling, de-noising, and modification. Effectively translating images between different domains requires semantic knowledge about pairwise embedding by exploiting natural correspondences. The correspondence between different domains can be categorized considering different aspects and problems. Such relations are naturally recognized by humans. For example, there is a natural relationship between an RGB image and its corresponding depth map, between edge-based representation and its real image correspondence, and between an aerial image and a map image. We explore this image-to-image translation task as a problem of translating image representation from one domain to the corresponding domain, given sufficient training data.

The image-to-image translation problem is related to either computer vision, where the mapping is from many to one, or computer graphics, where the mapping is from one to many. Despite the similar nature of these tasks, they have been tackled separately by [1–13]. However, in our approach, we tackled this in a unified framework. Moreover, the existing approaches are limited in performance in terms of generalization and accuracy because of imbalanced training strategies. Accurately understanding the scene and estimating the terrain from the available information are critical for collision and tip-over

avoidance in the field of robotics. Image-to-image translation plays a vital role in this; RGB images are translated to their corresponding depth maps by exploiting their natural correspondence. Although the current approaches are effective, they are less accurate and lack generalization. The proposed approach addresses these issues because of its balanced training strategy.

Recent advances in deep neural networks, specifically generative adversarial networks (GANs) with a generator and discriminator with a wide range of training data, are an efficient and powerful tool for high-quality image-to-image translation. This is because the objective of GAN is based on game theory, where the generator attempts to fool the discriminator by generating realistic images. The discriminator attempts to discriminate between real images and synthetic images generated by the generator. The generator and discriminator are trained until the discriminator fails to discriminate between the real and synthetic images. Although it has impressive performance in generalization and effective image-to-image translation, it has limitations such as stability and consistency in terms of training GAN-based image translation networks, which affect the accuracy of the network. To address the problem of inconsistency in training, we propose a consistent image-to-image translation network with improved accuracy and generalization. Further details are provided in Sections 3 and 4.

The remainder of this paper is organized as follows. Section 2 discusses the related work. Section 3 discusses the motivation and problem statement. Section 4 introduces the proposed consistent image-to-image translation network for accurate cross-domain image-to-image translation and discusses its training procedure. In Section 5, the experimental results are described, and they are analyzed. Section 6 presents the discussion and future scope. The last section concludes our study.

2. Related Work

Thus far, a significant amount of work has been done on the same and cross-domain image-to-image translation from the perspective of regression with convolutional neural networks (CNNs) as the basic platform for a wide variety of image prediction problems. Same-domain image-to-image translation has applications in domain adaptation [1–5,11,13] super-resolution [6], style transfer [7], and photo editing [8]. Cross-domain image-to-image translation has applications in data generation [9], data interpretation [10], and image completion [10,14,15]. The success of CNN-based approaches is because of the availability of a large amount of paired data with a natural correspondence. These approaches outperform state-of-the-art non-CNN approaches [16,17]. For the effective representation learning of image translation, several deep generative networks such as autoencoders (AEs) [18], variational AEs [19,20], GANs [9], moment-matching networks [21], pixel-CNN [22], and plug-and-play generative networks [23] have been proposed. Recently, many variants of deep AEs, and of GAN, have been proposed, including [19,20], LapGAN [24], DCGAN [25], WGAN [26], and the conditional generative adversarial network (cGAN) [10]. However, the combination of AE and GAN has shown the best performance in automated image translation [10].

Shrivastava et al. proposed image translation based on cGAN in [14]. They addressed the problem of performance degradation of a deep neural network for real data after training with synthetic data, by bridging the gap between the simulated and real images in the deep learning network. They translated synthetic images to real images using cGAN. Their study aimed to minimize the cost function based on the L_1 distance between the synthetic and real images along with the adversarial loss. This task of translating synthetic images to real images is simple with a simple loss function. However, the job becomes challenging when dealing with natural images. Minimizing the L_1 distance between the real and synthetic images is a challenge, both in terms of generalization and realistic translation.

Isola et al. proposed pix2pix cGAN in [10]. They used the exact correspondence of images in both domains for training the image translation network. Although it was effective at translating images from one domain to the corresponding domain, it had an imbalanced training strategy, thus lacking consistency. The pix2pix approach in [10] used the exact correspondence of pairs for image translation. In contrast, Yi, Zili, et al. proposed DualGAN [15] by exploiting dual learning to map the source and target images. DualGAN can learn image translation without exact correspondence

because its learning rule is based on the invertibility of the mapping while utilizing cyclic consistency loss. However, it fails to learn translation effectively when the corresponding objects in the target domain are not available [27].

Recently, in [28], BA-DualAEcomprising two AEs with individual latent spaces associated with a bidirectional regression network was proposed. It can translate images between different domains with an additional capability of image completion, proving its generality. However, it does not accurately recover local pixel intensity values. Furthermore, the association network in BA-DualAE translates the latent space of one domain to another using a fully connected network, which provides excellent generalization. However, it loses fine details in the input samples, resulting in low accuracy. BA-DualAE is effective for simple inputs where association does not affect the accuracy, but is limited in terms of scene images where accuracy is vital.

Conditional adversarial frameworks have shown limitations in generality and accuracy because of the inconsistency in the training strategy [10]. We attempt to address this issue by considering [10] as the basic platform. The proposed network shows improved performance in terms of accuracy, as discussed in Section 5.

3. Problem Statement

An image-to-image translation network based on a deep neural network (DNN) can be generalized because the DNN-based approaches are data driven and learn the representation of the data. In other words, an image-to-image translation network is not task specific and not limited by data variation. Despite its impressive performance, there are three main challenges: (1) accuracy, (2) slow convergence, and (3) consistency. To address these problems, an effective image translation network is required that can explore the natural correspondence between different domains. The goal of the image-to-image translation network is to translate images $\mathbf{x}$ of one domain x to images $\mathbf{y}'$ of the corresponding domain y as close as possible to target images $\mathbf{y}$ of domain y as follows.

$$G(\mathbf{x}) = \mathbf{y}' \text{ subject to } \mathbf{y}' \cong \mathbf{y} \tag{1}$$

where G is a generator that takes images $\mathbf{x}$ and generates the translated images $\mathbf{y}'$ subject to $\mathbf{y}' \cong \mathbf{y}$. In addition, an appropriate training methodology such as "consistency in training" is required for faster convergence with high accuracy.

To address the above-mentioned problems related to accuracy, convergence and consistency, we propose a consistent image-to-image translation network to translate images effectively from one domain to the required corresponding domain, as shown in Figure 1. The proposed network was inspired by cGAN [10], which takes advantage of the deep AE [18], volumetric CNN [8,18], and GAN [9]. The consistent image-to-image translation network is comprised of two cross-domain generators and a discriminator. The cross-domain generators translate the images from one domain to another. The real and generated fake pairs are input to the discriminator in three different configurations to achieve faster convergence and maintain consistency and high accuracy.

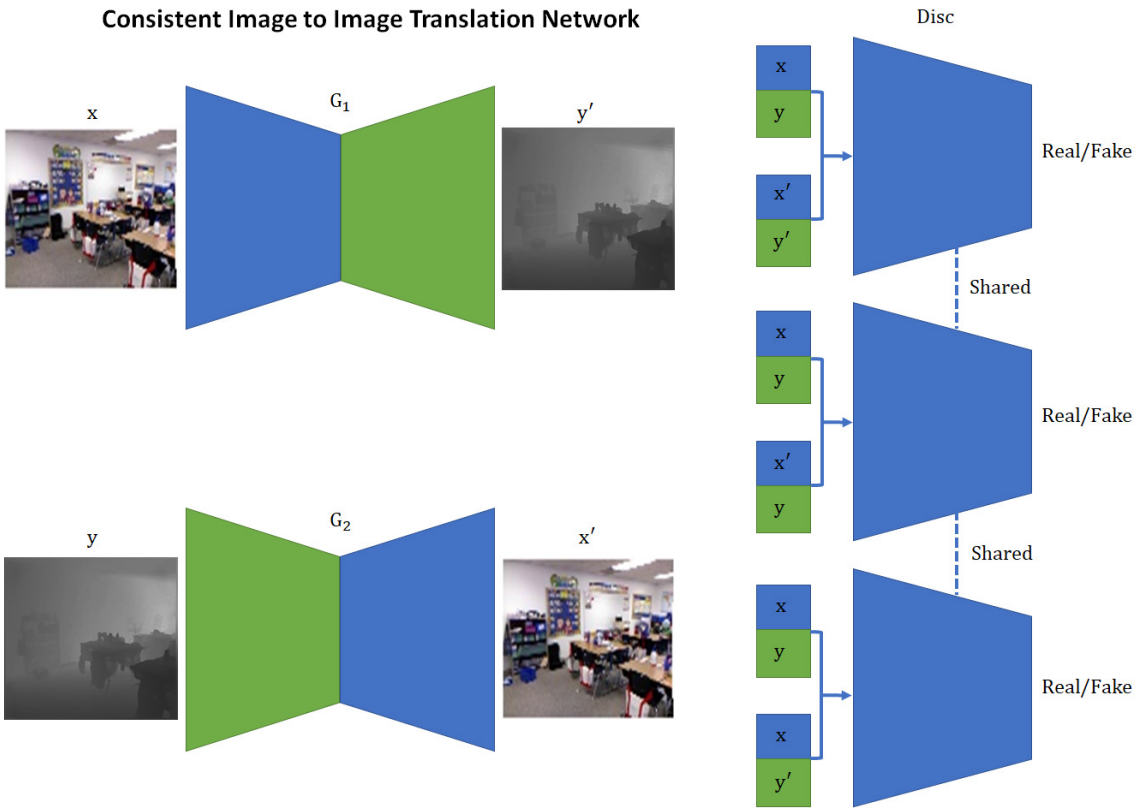

Figure 1. Consistent image-to-image translation network, where G_1 and G_2 are generators that translate images from domain x to domain y and from domain y to domain x, respectively. The discriminator takes the input in three different configurations, as shown on the right side of the figure.

4. Proposed Approach

To address the above-mentioned problems, the proposed consistent image-to-image translation network is comprised two cross-domain generators and a global discriminator. The cross-domain generators take images from their respective domains and generate images in the required cross-domains. The discriminator takes real and fake pairs in three different configurations. (1) In the first configuration, for the real pair, both inputs are real input images, and for the fake pair, both input images are generated by the cross-domain generators. (2) In the second configuration, the fake pair has one real image from domain x and one fake image from domain y. (3) In the third configuration, the fake pair consists of one real image from domain y and one fake image from domain x. The real pairs in all configurations are the same. Both cross-domain generators are trained simultaneously, and the training is end-to-end. The discriminator takes the input in the above-mentioned different configurations and guides both the cross-domain generators to generate realistic output images. The proposed approach provides stability in training and fast convergence while achieving a sufficient level of generality and high accuracy.

4.1. Network Architecture

The proposed consistent image-to-image translation network consists of two cross-domain generators and a discriminator. The first cross-domain generator G_1 takes the 2D input images $\mathbf{x}$ from domain x with a resolution of $[256 \times 256]$, and it translates them to their corresponding cross-domain representations as $G_1(\mathbf{x}) = \mathbf{y}'$ with the same resolution as that of the input images. In contrast, the second cross-domain generator G_2 translates the 2D input images $\mathbf{y}$ from domain y to their corresponding 2D images of domain x as $G_2(\mathbf{y}) = \mathbf{x}'$. We follow the architecture of cGAN [10], where both the cross-domain generators, G_1 and G_2, have the same configuration.

4.2. Training Details

To learn the parameters of the proposed network, we need to design an objective function that provides a more consistent model with high accuracy. In image-to-image translation tasks using a generative adversarial network for translating images from one domain to another domain, consistency plays a significant role in the successful training of the generator and the discriminator and for generating more realistic and accurate results. The objective function in the proposed network is designed in such a way that it controls the learning of the generator and discriminator at an equal pace to ensure consistency in training. In image translation tasks, the discriminator takes a pair of real images and classifies them as a real pair in the first step. Next, the discriminator takes a fake pair and classifies them as fake. The fake pair is the output of the generator; however, in the existing cGAN-based approach [10], the fake pair contains one real image and one generated image. Hence, it results in imbalanced training of the discriminator, which affects the training of the generator in producing more realistic images. In the proposed training approach, we provide every possible combination of the inputs as a fake pair to the discriminator by adding one more generator in the reverse direction.

The training procedure of the proposed consistent image-to-image translation network is given below. The objective function of the discriminator is as follows.

$$L_D = \gamma_d(D(x,y) + 1 - D((G_1(x), G_2(y))) + D(x,y) + 1 - D((G_1(x), y)) + D(x,y) + 1 - D((x, G_2(y)))) \quad (2)$$

where γ_d is the learning rate and (x, y) is the real pair input to the discriminator. $(G_1(x), G_2(y))$ represents a fake pair where both images are generated by G_1 and G_2, respectively. $(G_1(x), y)$ is a fake pair where $G_1(x)$ is the domain x fake image generated by G_1 and y is the real image from domain y. Similarly, $(x, G_2(y))$ is a fake pair where x is the real image from domain x and $G_2(y)$ is the domain y fake image generated by G_2.

The objective functions of the cross-domain generators are based on the weighted mean squared error loss along with the conventional generative adversarial loss. The weighted mean squared loss is used to quantify the differences between the real data and the corresponding model outputs along with providing stability to the generative adversarial network learning.

$$L_{G1} = \gamma_g(1 - D((G_1(x), G_2(y))) + 1 - D((G_1(x), y)) + 1 - D((x, G_2(y)))) + \gamma||x - x'||_2^2 \quad (3)$$

$$L_{G2} = \gamma_g(1 - D((G_1(x), G_2(y))) + 1 - D((G_1(x), y)) + 1 - D((x, G_2(y)))) + \gamma||y - y'||_2^2 \quad (4)$$

where γ_g and γ are the learning rates for the GAN loss and the mean squared error loss, respectively, and $||x - x'||_2^2$ and $||y - y'||_2^2$ represent the mean squared error losses of generators G_1 and G_2, respectively. Finally, L_{G1} and L_{G2} are the total losses of generators G_1 and G_2, respectively. To optimize the parameters of the proposed framework, we used the ADAM optimizer [29] with $\beta = 0.5$.

5. Results

The main purpose of our approach was to perform robust and accurate translation between different domain images while maintaining consistency in training and fast convergence. For this analysis, we used two different datasets, NYU [30] and Cityscapes [31], and evaluated the approach qualitatively and quantitatively by performing different experiments.

The NYU dataset is comprised of approximately 50,000 pairs of RGB and their corresponding depth images. However, in the experiments, we selected 20,000 samples and then randomly selected 5% testing samples from them. The remaining samples were used as training samples. The Cityscapes dataset is comprised of approximately 2957 pairs of RGB and their corresponding label samples for training. We randomly selected 500 sample pairs for testing.

We trained the proposed consistent image-to-image translation network to learn the translation between different domain samples by following the training algorithm as discussed above.

We analyzed the experimental results qualitatively and quantitatively by defining the mean squared error (MSE) and structural similarity index (SSIM) from the ground truth for the translated images, as shown in Tables 1 and 2. The consistent image-to-image translation network was implemented by using TensorFlow 1.13, an open-source deep learning framework, on the GPU-based PC, which was comprised of an Intel(R) Core i9-9940X CPU, 132.0 GB RAM, and four NVIDIA GeForce RTX 2080 Ti graphics cards.

Table 1. MSE and SSIM based on reconstructing input test samples using the NYU dataset [30].

Approach	MSE per Pixel	MSE per Image	SSIM per Pixel	SSIM per Image
BA-DualAE-based Approach [28]	0.0010937	71.6767	8.7586^{-06}	0.574
cGAN-based Approach [10]	0.0002298	15.0585	1.2541^{-05}	0.821
Proposed Approach	0.0001738	11.3912	1.3352^{-05}	0.875

Table 2. MSE and SSIM based on reconstructing input test samples using the Cityscapes dataset [31].

Approach	MSE per Pixel	MSE per Image	SSIM per Pixel	SSIM per Image
BA-DualAE-based Approach [28]	0.0009063	59.4014	6.7803^{-6}	0.4443
cGAN-based Approach [10]	0.0005462	35.7950	9.5650^{-6}	0.6268
Proposed Approach	0.0004449	29.1625	1.1113^{-5}	0.7283

5.1. Qualitative Analysis

After training the proposed consistent image-to-image translation network for translating images from one domain to another by using different datasets, we evaluated its performance in terms of realistic translation between different domains. In the first experiment, we used the NYU dataset [30] for training and testing, and in the second experiment, we used the Cityscapes dataset. The performance of the proposed approach was evaluated qualitatively in comparison with the cGAN-based approach using the test datasets. The qualitative analysis is shown in Figures 2–6.

5.1.1. Translation from Domain x to Domain y

In the first experiment, we performed a comparative qualitative analysis of translating images from domain x to domain y. Domain x contained real natural scene images, and domain y contained the depth or labels of the scenes in domain x. We used two different datasets in the analysis. The first analysis was based on the NYU dataset [30], as shown in Figure 2, and the second on the Cityscapes dataset, as shown in Figure 3. In the first analysis, we first translated images from domain x to domain y and then evaluated the translated results based on the L^1 norm between the ground truth images and the translated images. Figure 2 summarizes the overall comparative results for the NYU dataset. The first row shows the real natural scene images used as input; the second shows the ground truth images; and the third shows the translated images from the proposed consistent image-to-image translation network. To show the effectiveness of the proposed approach in translating images from one domain to the corresponding cross-domain, we calculated the L^1 norm between the ground truth images and the translated images by using the proposed approach. The resulting images were considered as the error between the ground truth images and the generated images, shown in the fourth row of Figure 2. The fifth row shows the images generated by the cGAN-based approach, and the last row shows the error between the ground truth images and the generated images by the cGAN-based approach.

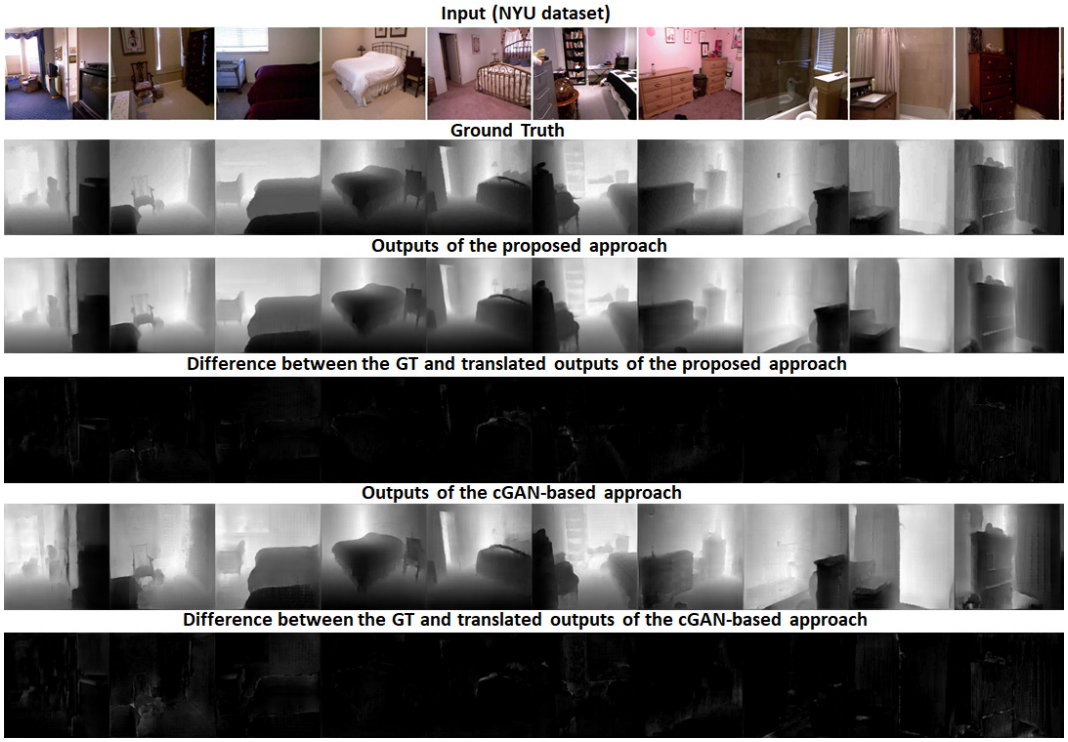

Figure 2. Input test samples in domain x are translated to domain y samples by the proposed approach and the cGAN-based approach [10]. The difference between the ground truth and generated images of the proposed approach and the cGAN-based approach are compared.

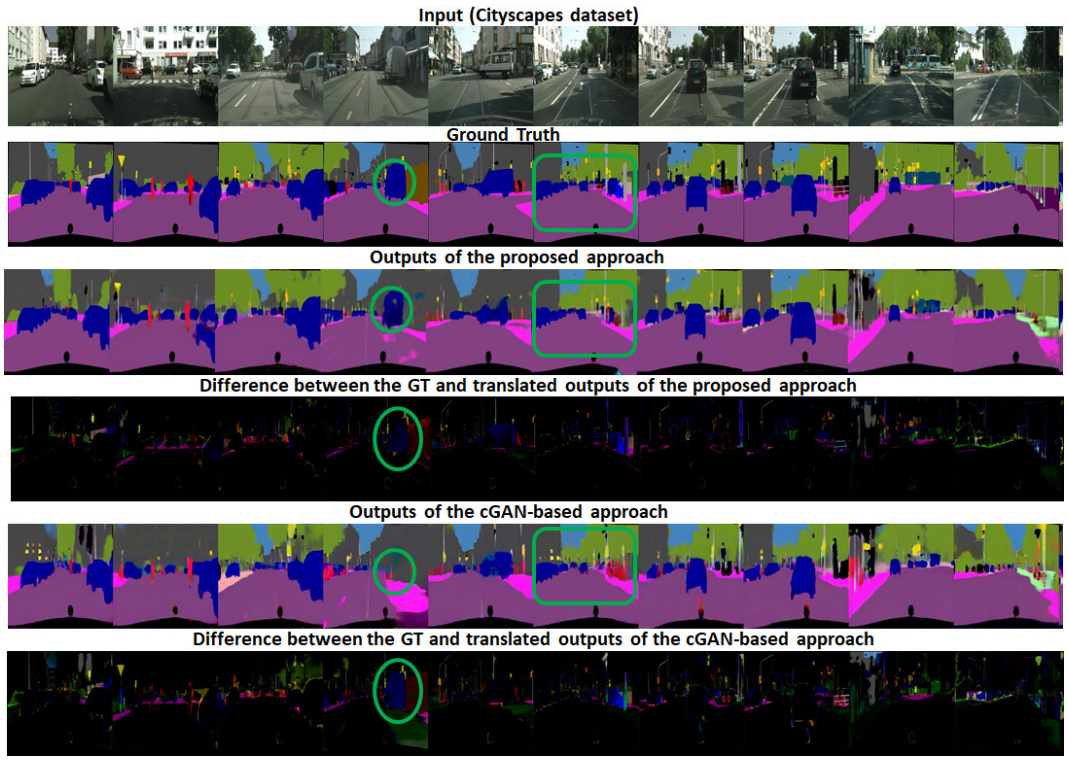

Figure 3. Input test samples in domain x are translated to domain y samples by the proposed approach and cGAN-based approach [10]. Comparative analysis was performed for the difference between the ground truth and images generated by the proposed approach and the cGAN-based approach.

The lower the L^1 norm was, the more accurate and realistic the image translation was. Here, the L^1 norm was the pixel-wise difference between the ground truth images and the generated images from the proposed approach and the cGAN-based approach. We calculated the L^1 norm to show the effectiveness of the proposed approach in terms of per-pixel loss. From the fourth row, it can be seen that the difference between the translated images and their corresponding ground truth images was small compared with the cGAN-based approach in the last row. The shade of the objects is more prominent in the last row compared with the fourth row (the proposed approach) because of the high per-pixel loss. This showed that the proposed approach had more realistic translation compared with the cGAN-based approach. When directly comparing the outputs of the proposed approach and cGAN-based approach (the third and fifth rows), it was difficult to see the difference between them. However, the L^1 norm highlighted the error regions, thus providing a meaningful way for analyzing the performance of the proposed approach. In the second analysis, we used the Cityscapes dataset and followed the same procedure to evaluate the effectiveness of the proposed approach. The overall analysis is shown in Figure 3. The input images are shown in the first row, which were translated by the proposed approach and the cGAN-based approach, shown in the third and fifth rows, respectively. The L^1 distance between the outputs and their corresponding ground truth images is shown in the fourth and last rows, respectively. From the translated images, we can observe that the fine details in the images generated by the cGAN-based approach were not well represented, compared with the proposed approach. In Figure 3, the electric poles in the green box were not well translated by the cGAN-based approach, whereas the output of the proposed approach in the third row was closer to the ground truth image in the second row. Furthermore, the "van," which is indicated by the green circle in the second row, was missing in the image translated by the cGAN-based approach, as shown in the sixth row; it was well translated by the proposed approach, shown in the third row. Moreover, the green ovals in the fourth and last rows represent the high-error regions, where the error was more prominent in the output of the cGAN-based approach compared with the proposed approach. This clearly showed the advantage of the proposed approach over the cGAN-based approach.

5.1.2. Translation from Domain y to Domain x

In the second experiment, we performed a comparative qualitative analysis for translating images from domain y to domain x using the NYU dataset [30]. First, we translated images from domain y to domain x and then evaluated the translated results based on the L^1 norm between the ground truth and translated images. Figure 4 summarizes the overall comparative results based on the NYU dataset. The first row shows the input domain y images; the second shows the ground truth images; and the third shows the images translated by the proposed approach. The error between the ground truth and the generated images is shown in the fourth row of Figure 4. The fifth row shows the images generated by the cGAN-based approach, and the last shows the error between the ground truth and the images generated by the cGAN-based approach. We followed the same procedure discussed in the previous section to evaluate the effectiveness of the proposed approach and the cGAN-based approach for translating images to the corresponding cross-domain. We observed that the proposed approach outperformed the cGAN-based approach in translating images from domain y to domain x.

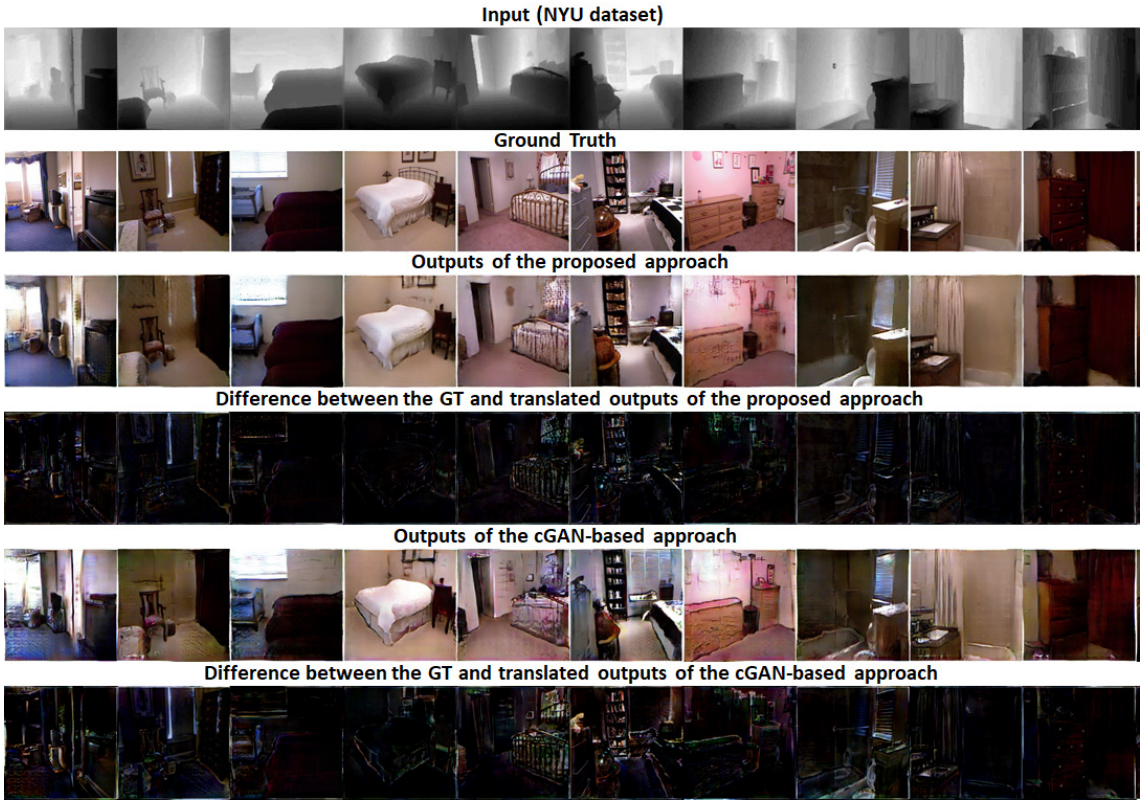

Figure 4. Input test samples in domain y are translated to domain x samples by the proposed approach and cGAN-based approach [10]. The comparative analysis was performed for the difference between the ground truth and images generated by the proposed approach and the cGAN-based approach.

5.1.3. Analyzing the High Erroneous Regions

Figures 2–4 compare the proposed approach with the cGAN-based approach. However, to further elaborate the comparative analysis, we highlighted the high-error regions. For this, we used a threshold value for the error output and marked the high-error regions in red, as shown in the fourth and last rows of Figures 5 and 6. Figure 5 shows the comparative analysis of the NYU dataset and Figure 6 of the Cityscapes dataset. After applying the threshold to the erroneous output, we analyzed the results. The proposed approach had fewer erroneous regions than the cGAN-based approach for both the NYU and Cityscapes datasets. For example, the high-error regions in Figure 5 inside the green boxes from the proposed approach were almost negligible, compared with the high-error regions inside the yellow boxes from the cGAN-based approach. Furthermore, for the proposed approach, the error regions were mostly lying on the edges where the gradient was high. This was understandable because the nature of the convolution operation that learned the correlation in the pixels in the input space resulted in the smoothing of the boundaries of the high gradient regions. The output of the cGAN-based approach was affected by the boundary regions and also had erroneous regions in the low gradient regions.

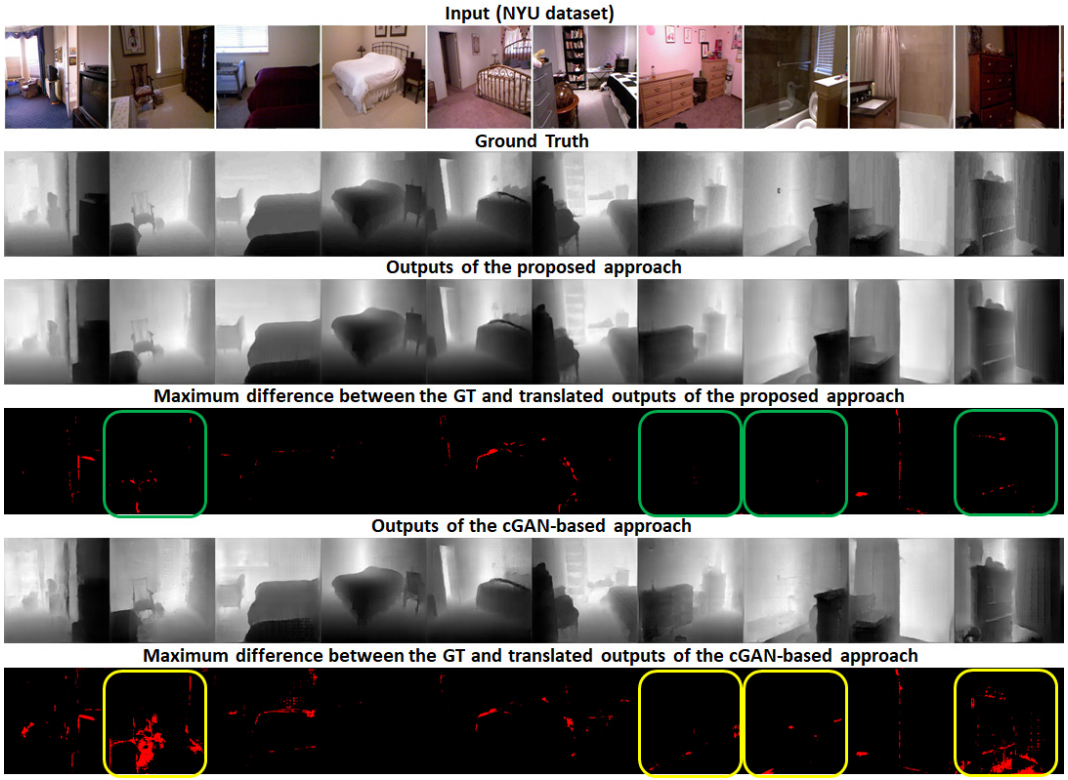

Figure 5. Comparative analysis of the proposed approach with cGAN [10] with emphasis on the high-error regions using the NYU dataset.

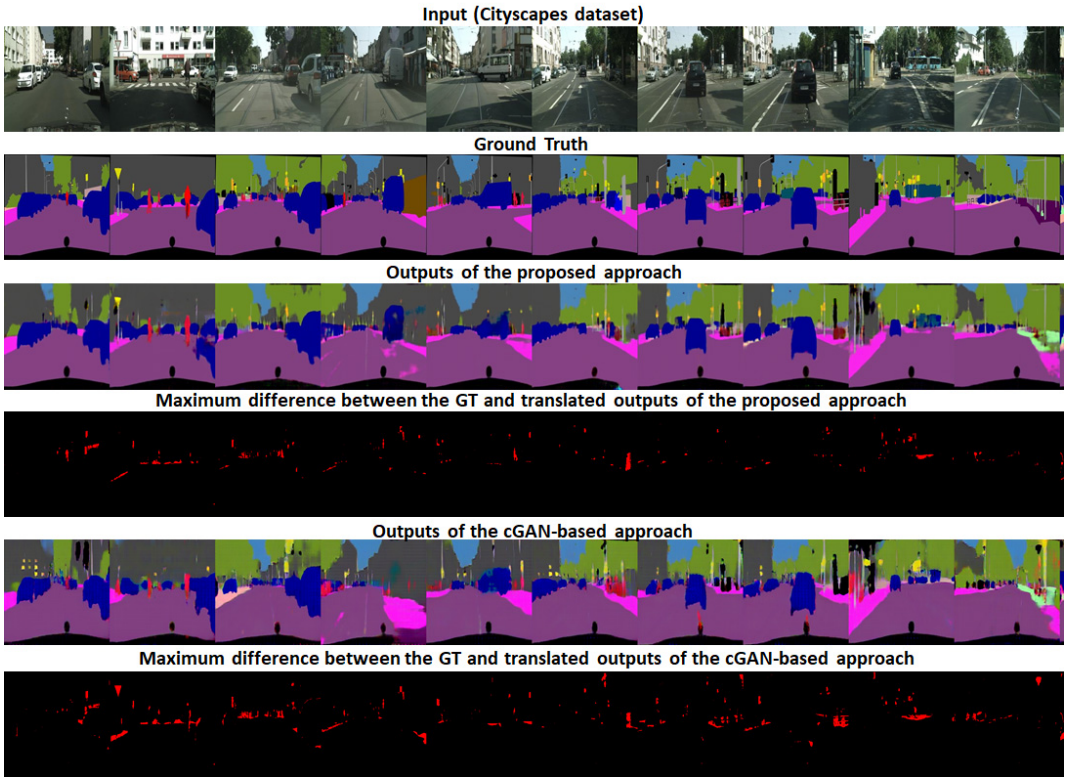

Figure 6. Comparative analysis of the proposed approach with cGAN [10] with emphasis on the high-error regions using the Cityscapes dataset.

5.2. Quantitative Analysis

The qualitative results discussed above showed the effectiveness of the proposed approach for realistically translating images from one domain to their corresponding cross-domain, compared with the cGAN-based approach. However, for insight into the analysis of the proposed approach, we performed a quantitative analysis in this section. We used two different quantitative measures, the mean squared error (MSE) and SSIM, for the comparative analysis [10]. First, we randomly selected two batches of input test samples from the NYU [30] and Cityscapes [31] datasets and translated these samples to their corresponding cross-domains by using the proposed approach, the cGAN-based approach [10], and the BA-DualAE-based approach [28]. The sample-wise comparative analysis for the NYU and Cityscapes datasets in terms of the MSE is shown in Figure 7a,b, respectively. The green plot shows the error of the proposed approach, and the blue and cyan plots show those of the cGAN- and the BA-DualAE-based approaches, respectively. We could observe that 70% of the samples translated by the proposed approach from the NYU dataset had fewer errors than the cGAN-based approach, and more than 80% of the translated samples had fewer errors than the BA-DualAE-based approach. For the Cityscapes dataset, all samples generated by the proposed approach had a lower MSE than the cGAN- and the BA-DualAE-based approaches. Furthermore, the per-pixel and per-image MSEs between the ground truth and the generated samples from the proposed approach, the cGAN-based approach, and the BA-DualAE-based approach are listed in Tables 1 and 2. Tables 1 and 2 summarize the comparative analysis for the NYU dataset and the Cityscapes dataset, respectively. This quantitative analysis showed that the proposed approach outperformed the cGAN- and the BA-DualAE-based approaches by more than 26%. Furthermore, we also analyzed the structure of the images generated by the proposed approach, the cGAN-based approach, and the BA-DualAE-based approach; for this, we used SSIM. SSIM showed the structural similarity between the data generated by the network and the corresponding ground truth samples. The higher the SSIM was, the better the results were. The per-pixel and per-image SSIM results for the NYU and Cityscapes datasets are listed in Tables 1 and 2, respectively. The per-pixel and per-image SSIM values were higher for the proposed approach than the cGAN- and the BA-DualAE-based approaches for both datasets, showing that the proposed approach outperformed the cGAN- and the BA-DualAE-based approaches.

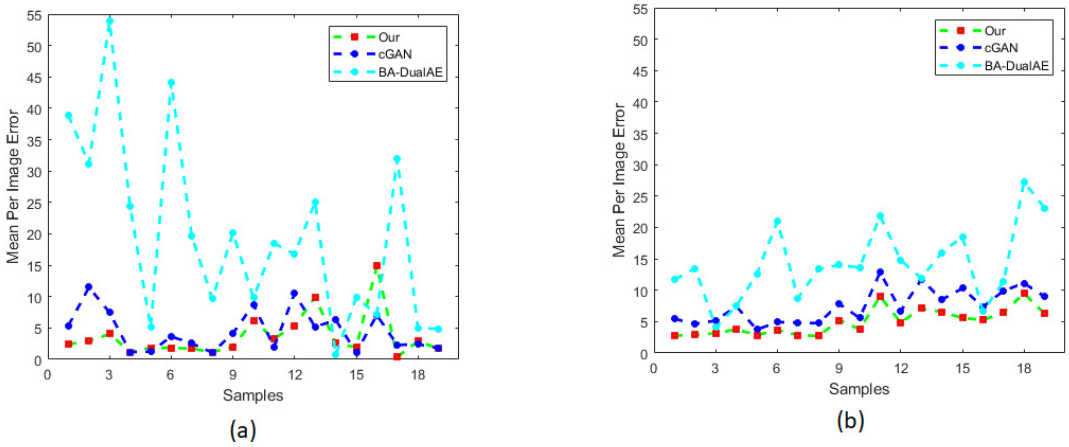

Figure 7. Per-image MSE of randomly selected input test samples from the (**a**) NYU dataset and (**b**) Cityscapes dataset. The green plot shows the MSE of each input sample, and the blue and cyan plots represent the MSEs from cGAN [10] and BA-DualAE [28], respectively.

6. Discussion

Image-to-image translation can be broadly categorized as explicit and implicit image translation tasks based on its application. The problem of implicit image translation is multimodal, and it

has applications in domain adaptation, super-resolution, imaging, and photo editing. Thus far, it has shown impressive performance in domain adaptation [1–5,11,13], super-resolution [6], style transfer [7], and photo editing [8]. In contrast, the explicit image translation was unimodal, and the image from one domain was explicitly translated into its cross-domain. Recent work on explicit image translation was done in [2,10,11]. The current approaches for explicit image translation were based on GANs. GAN is comprised of two networks, the discriminator and generator. Its learning was based on game theory, where the generator attempted to fool the discriminator by generating realistic images. The discriminator attempted to discriminate between real images and synthetic images generated by the generator. The generator and discriminator were trained until the discriminator failed to discriminate between the real and synthetic images. This training mechanism was based on the adversarial objective, which was sensitive in terms of stability, which resulted either in failed training or low accuracy. Several learning tricks and variants of the GAN have been proposed to overcome the stability issues of GAN, such as LapGAN [24], DCGAN [25], WGAN [26], cGAN [10], and EBGAN [32]. This study focused mainly on minimizing the effect of adversarial loss by embedding the AE loss either in the generator or discriminator part of the network. However, in our study, we selected cGAN [10] as the first step of our approach and carefully analyzed the learning mechanism. We found that the learning of the cGAN [10] was inconsistent, where the discriminator took a pair of real and fake images. The fake pair in the cGAN [10] had one real input image and one generated image, and the real pair had both real images. The discriminator discriminated between real and fake pairs; however, in this scenario, the task of the discriminator was easy, and it converged faster than the generator. Hence, it resulted in low generator output accuracy. We proposed a consistent image-to-image translation network to address the problem of inconsistency in cGAN [10]. The consistent image-to-image translation network was comprised of two cross-domain generators and one discriminator. The cross-domain generators translated images from their respective domains to the corresponding cross-domains. The discriminator took the input of real and fake pairs in three different configurations and guided the cross-domain generators to generate accurate results. In this way, the proposed approach provided stability in training and fast convergence, while achieving a sufficient generality and high accuracy.

To further elaborate the stability and accuracy related to the existing cGAN-based approach in terms of imbalanced and inconsistent training, we considered an example given in Figure 8. The first image is the input to the image-to-image translation network; the second is the ground truth representation; and the last is the output of the cGAN-based approach. The training mechanism was imbalanced because the discriminator took the real and fake pairs in an imbalanced manner, resulting in unstable training. This made the discriminator converge faster than the generator. The fast learning of the discriminator inhibited the realistic output by the generator, which affected the accuracy. In Figure 8, the green boxes in the input and the corresponding ground-truth images show a "van." The last image is the output of the cGAN-based approach where the "van" was not translated in the output image. Similarly, the red boxes in the input and the corresponding ground-truth images have different object labels, but the translated image from the cGAN-based approach showed the group of "people".

To address the challenges of low accuracy due to an imbalanced and inconsistent training strategy, we modified cGAN with a balanced and consistent training strategy, where the discriminator consumed every possible combination of the fake pairs. We then evaluated the effectiveness of the modified training strategy in terms of realistic image translation, as shown in Figure 9. In Figure 9, we can observe that the "van" was translated in the output image, which was missing in Figure 8 with an imbalanced training strategy. Moreover, the red box objects in Figure 9 were also translated reasonably well, compared with the imbalanced cGAN in Figure 8. This showed the effectiveness of the balanced training strategy in terms of realistic image-to-image translation.

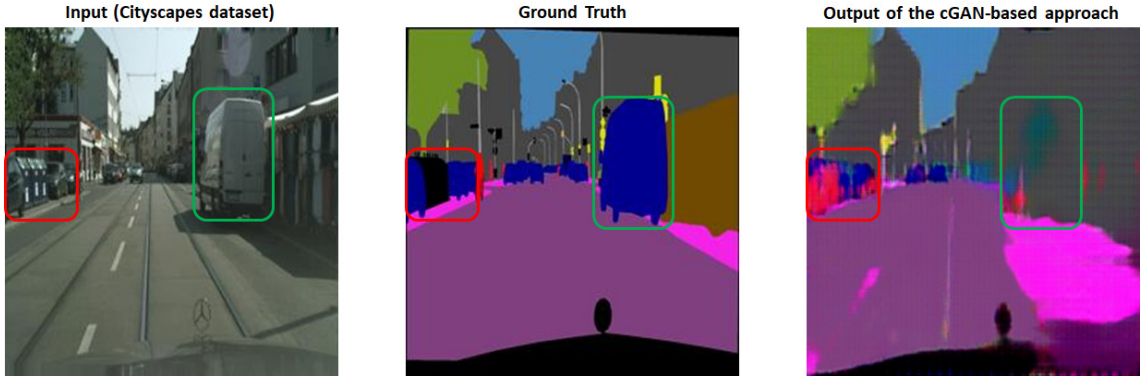

Figure 8. Evaluating the stability and accuracy of the GAN, where imbalanced training results in instability that affects the accuracy.

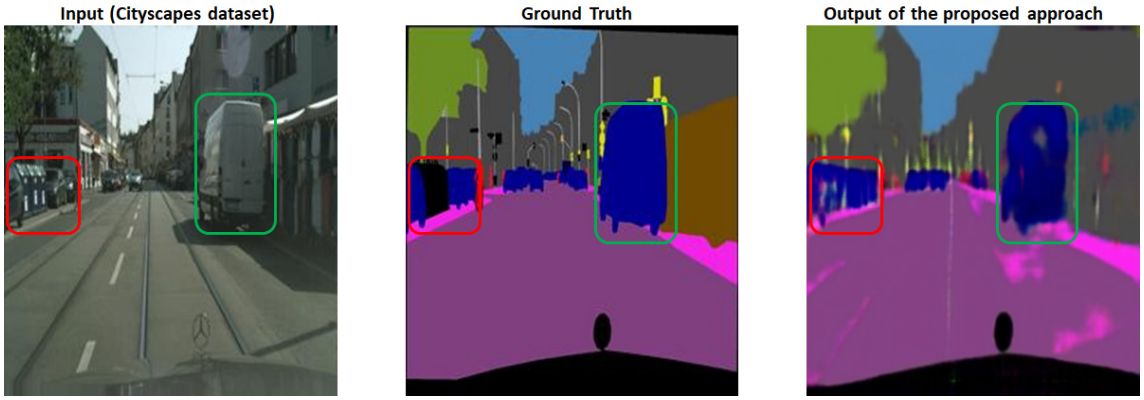

Figure 9. Evaluating the stability and accuracy of the consistent image-to-image translation network, where the balanced training results in stable learning that generates more accurate results.

7. Conclusions

We presented a consistent bidirectional image-to-image translation network, called the consistent image-to-image translation network. As we demonstrated, the proposed network was highly robust for generating realistic images with a high degree of generality with high accuracy. For the MSE, which represented a rough measure of robustness, the network was 26% better than the conventional cGAN-based approach. Furthermore, the balanced training strategy of the proposed network, where the discriminator took three different configurations of fake input pairs, ensured consistency and accuracy. The experimental results showed the effectiveness of the proposed network for generating realistic images in the required domain. In our future work, we will extend the image translation with further analysis for image modification, image style transfer, and accurate terrain shape estimation.

Author Contributions: Conceptualization, N.U.I., S.L., and J.P.; methodology, N.U.I.; software, N.U.I.; validation, N.U.I.; formal analysis, N.U.I. and J.P.; investigation, N.U.I.; resources, N.U.I.; data curation, N.U.I.; writing, original draft preparation, N.U.I.; writing, review and editing, N.U.I. and J.P.; visualization, N.U.I.; supervision, J.P.; project administration, J.P.; funding acquisition, J.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly supported by the Basic Science Research Programs (NRF-2019R1A6A1A09031717 and NRF-2018R1D1A1B07049270) through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, and research funds of Jeonbuk National University in 2016.

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

References

1. Taigman, Y.; Polyak, A.; Wolf, L. Unsupervised cross-domain image generation. *arXiv* **2016**, arXiv:1611.02200.
2. 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, Sydney, Australia, 6–11 August 2017.
3. Liu, M.Y.; Breuel, T.; Kautz, J. Unsupervised image-to-image translation networks. In Proceedings of the Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
4. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
5. 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 Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
6. 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. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
7. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
8. Shu, Z.; Yumer, E.; Hadap, S.; Sunkavalli, K.; Shechtman, E.; Samaras, D. Neural face editing with intrinsic image disentangling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
9. Goodfellow, Z.; Welling, M.; Cortes, C.; Lawrence, N.D.; Weinberger, K.Q. Generative adversarial nets. In Proceedings of the Annual Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014.
10. 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.
11. Islam, N.U.; Lee, S. Cross Domain Image Transformation Using Effective Latent Space Association. In Proceedings of the 15th International Conference IAS-15, Baden-Baden, Germany, 11–14 June 2018.
12. Islam, N.U.; Lee, S. Interpretation of deep CNN based on learning feature reconstruction with feedback weights. *IEEE Access* **2019**, *7*, 25195–25208. [CrossRef]
13. Islam, N.U.; Lee, S. Learning Typical 3D Representation from a Single 2D Correspondence using 2D-3D Transformation Network. In Proceedings of the International Conference on Ubiquitous Information Management and Communication, Phuket, Thailand, 4–6 January 2019.
14. Shrivastava, A.; Pfister, T.; Tuzel, O.; Susskind, J.; Wang, W.; Webb, R. Learning from simulated and unsupervised images through adversarial training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
15. 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, Honolulu, HI, USA, 21–26 July 2017.
16. Cheng, Z.; Yang, Q.; Sheng, B. Deep colorization. In Proceedings of the IEEE International Conference on Computer Vision, Boston, MA, USA, 7–12 June 2015.
17. 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 Trans. Graph. (TOG)* **2016**, *35*, 110. [CrossRef]
18. Baldi, P. Autoencoders, unsupervised learning, and deep architectures. In Proceedings of the ICML Workshop on Unsupervised and Transfer Learning, Edinburgh, UK, 26 June 26–1 July 2012.
19. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
20. Rezende, D.J.; Mohamed, S.; Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv* **2014**, arXiv:1401.4082.

21. Li, Y.; Swersky, K.; Zemel R. Generative moment matching networks. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015.
22. Van den Oord, A.; Kalchbrenner, N.; Espeholt, L.; Vinyals, O.; Graves, A. Conditional image generation with pixelcnn decoders. In Proceedings of the Annual Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016.
23. Nguyen, A.; Clune, J.; Bengio, Y.; Dosovitskiy, A.; Yosinski, J. Plug and play generative networks: Conditional iterative generation of images in latent space. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
24. Denton, E.L.; Chintala, S.; Fergus, R. Deep generative image models using a laplacian pyramid of adversarial networks. In Proceedings of the Annual Conference on Neural Information Processing Systems, Montréal, QC, Canada, 7–12 December 2015.
25. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
26. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein gan. *arXiv* **2017**, arXiv:1701.07875.
27. Zhou, S.; Xiao, T.; Yang, Y.; Feng, D.; He, Q.; He W. Genegan: Learning object transfiguration and attribute subspace from unpaired data. *arXiv* **2017**, arXiv:1705.04932.
28. Lee, S.; Islam, N.U. Robust Image Translation and Completion Based on Dual Auto-Encoder with Bidirectional Latent Space Regression. *IEEE Access* **2019**, *7*, 58695–58703. [CrossRef]
29. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
30. Silberman, N.; Fergus, R. Indoor scene segmentation using a structured light sensor. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops, Barcelona, Spain, 6–13 November 2011.
31. 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, 27–30 June 2016.
32. Zhao, J.; Mathieu, M.; LeCun, Y. Energy-based generative adversarial network. *arXiv* **2016**, arXiv:1609.03126.