2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

In2I : Unsupervised Multi-Image-to-Image Translation Using Generative Adversarial Networks

Pramuditha Perera, Mahdi Abavisani, and Vishal M. Patel
Rutgers University, Department of Electrical and Computer Engineering
94 Brett Road, Piscataway, NJ 08854, USA
pramuditha.perera@rutgers.edu, mahdi.abavisani@rutgers.edu, vishal.m.patel@rutgers.edu

***Abstract*—In unsupervised image-to-image translation, the goal is to learn the mapping between an input image and an output image using a set of unpaired training images. In this paper, we propose an extension of the unsupervised image-to-image translation problem to multiple input setting. Given a set of paired images from multiple modalities, a transformation is learned to translate the input into a specified domain. For this purpose, we introduce a Generative Adversarial Network (GAN) based framework along with a multi-modal generator structure and a new loss term, *latent consistency loss*. Through various experiments we show that leveraging multiple inputs generally improves the visual quality of the translated images. Moreover, we show that the proposed method outperforms current state-of-the-art unsupervised image-to-image translation methods.**

I. Introduction

The problem of unsupervised image-to-image translation has made promising strides with the advent of Generative Adversarial Networks (GAN) [6] in recent years. Given an input from a particular domain, the goal of image-to-image translation is to transform the input onto a specified second domain. Recent works in image-to-image translation has successfully learned this transformation across various tasks including satellite images to map images, night images to day images, greyscale images to color images etc. [30], [13], [12] [27].

In this work, we propose an extension of the original problem from a single input image to multiple input images, called multi-image-to-image translation $(I^n 2I)$. Given semantically related multiple images across n number of different domains, the goal of $I^n 2I$ is to produce the corresponding image in a specified domain. For example, the traditional problem of translating a greyscale image onto the RGB domain can be extended into an $I^n 2I$ problem by providing the near infrared (NIR) image of the same scene as an additional input. Now, the objective would be to use information present in greyscale and NIR domains to produce the corresponding output in the RGB domain. We study the problem of $I^n 2I$ in the more generic unsupervised setting and provide initial direction to solve the problem.

Image-to-image translation is a challenging problem. For a given input, there exists multiple possible representations in the specified second domain. Having multiple inputs from different image modalities reduces this ambiguity due to the presence of complimentary information. Therefore, as we show later in experimental results section, leveraging multiple input images leads to an output of higher perceptual quality. Multiple input modalities can be incorporated naively by concatenating all available modalities as channels and feeding into an existing image-to-image translation algorithm. However, when the input modalities are from incompatible domains, such a fusion scheme results in incoherent reconstructions as will be shown later. Therefore, we argue that unsupervised multi-image-to-image translation should be treated as a unique problem. In this work our main contributions are three-fold: 1) We introduce the problem of unsupervised multi-image-to-image translation. We show that late fusion of multiple modalities results in a better quality output in the desired domain. 2) A GAN-based scheme is proposed to combine information of multiple modalities to produce the corresponding output from the desired domain. We introduce a new *latent consistency loss* term, into the objective function. 3) We propose a generalization to the GAN generator network by introducing a multi-modal generator structure.

II. Related Work

To the best of our knowledge, $I^n 2I$ problem has not been previously addressed in the literature. In this section, we outline previous work related to the proposed method.

Generative Adversarial Networks (GANs). The fundamental idea behind GAN introduced in [6],[20] is to use two competing Fully Convolutional Networks (FCN [21]), the generator and the discriminator, for generative tasks. Learning objective of this problem is collectively called as the *adversarial loss* [30]. Many applications have since employed GANs for various image generation tasks with success [18], [9], [11], [8], [25], [29], [28], [14], [22], [13], [24], [4].

Unpaired image-to-image translation. Several recent methods have addressed the unsupervised image-to-image translation task when the input is a single image. Here, unlike in the supervised setting, paired samples across the two domains do not exist. In [30], image-to-image translation problem is tackled by having two generators and discriminators, one for each domain. In addition to the adversarial loss, a cycle consistency constraint is added to ensure that the semantic information is preserved in the translation. A similar rationale is adopted in DualGAN [27] which has been developed independently of CycleGAN. In [12], the CoGAN framework [13] was extended using GANs and variational autoencoders with the assumption of a common latent space between the domains.

978-1-5386-3788-3/18/$31.00 ©2018 IEEE

Image fusion. Although image fusion [15] operates on multiple input images, we note that our task is very different from image fusion since the former does not involve a domain translation process. In image fusion tasks, multiple input modalities are combined in an informative latent space. This space is usually found by a derived multi-resolution transformation such as wavelets [17]. In [16] operating on deep networks, a latent space is used to re-generate outputs of multiple modalities. Motivated by this technique, we fuse mid-level deep features from each input domain in the proposed generator FCN.

III. Proposed Method

Notation. In this paper, we use the following notations. Source domain and target domain are denoted by S and T, respectively. The latent space is denoted by Z. In the presence of multiple source domains, the set of source domains $\{S_1, \ldots, S_n\}$ are denoted collectively as $\mathcal{S}$. A data sample drawn from an arbitrary domain X is denoted as x. The transformation between domains X and Y is denoted by the function $f_{X \to Y}$. The transformation between the domains X and the latent space Z is denoted by $h_{X \to Z}$.

Overview. In conventional image-to-image translation, the objective is to translate images from an original domain S to a target domain T using a learned transformation $f_{S \to T}(.)$. In the supervised setting of the problem, a set of image pairs $\{(s_1, t_1), (s_2, t_2), \ldots, (s_p, t_p)\}$ are given, where $s_i \in S$ and $t_i \in T$ are paired images from the two domains. Image-to-image translation task is less challenging in this scenario since the desired output for a given input is known ahead of time.

Similar to the supervised version of the problem, images from both target and source domains are provided in the unsupervised image-to-image translation problem. However, in this case, provided images of the two domains are not paired. In other words, for a given source image s_i, the corresponding ground truth image t_i is not provided. In the absence of image pairs from both domains, it is not possible to optimize over a distance between the estimated output and the target. One possible option is to introduce an *adversarial loss* to facilitate reward if the generated image is from the same domain as the target domain. However, having an adversarial loss alone does not guarantee that the generated image will share semantics with the input. Therefore, to successfully solve this problem, additional constraints need to be imposed.

In [30], such a solution is sought by enforcing the cycle consistency property. Here, an inverse transformation $f_{T \to S}(.)$ is learned along with $f_{S \to T}(.)$. Then, the cycle consistency ensures that the learned transformation yields a good approximation of the input s_i by comparing s_i with $f_{T \to S}(f_{S \to T}(s_i))$. We develop our method based on the foundations of CycleGAN proposed in [30]. Here, we briefly review the CycleGAN method and we will draw differences between CycleGAN and our method in succeeding sections. CycleGAN as shown in Figure 1 (a) (top), contains a forward transformation from source domain to target domain and a reverse transformation from target to source. Two discriminators D_S and D_T are used to asses whether a given input belongs to source or target, respectively.

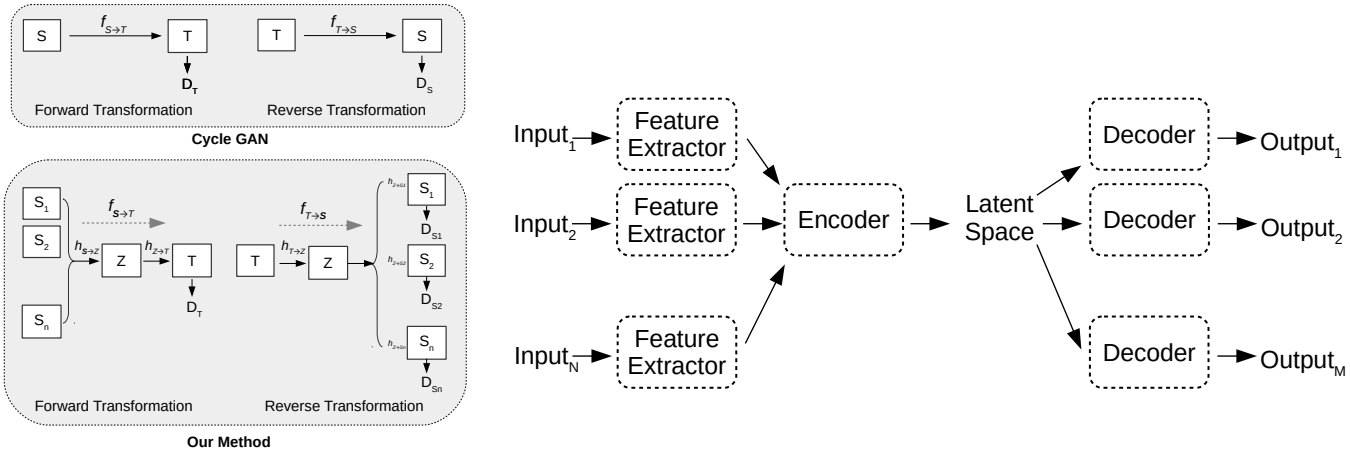
Multimodal Generator. The I^n2I problem accepts n inputs and translates them into a single output. Therefore, in contrast to CycleGAN, the proposed method deals with multiple inputs in the forward transformation and multiple outputs in the reverse transformation. In order to facilitate this operation, we propose a generalization of the generator structure for multiple inputs and outputs. The generic structure of the proposed generator is shown in Figure 1(b). In general, it is possible for the generator to have N inputs and M outputs. The generator treats each input modality independently and extracts features and fuses them prior to feeding them to the encoder. The encoder maps resultant features to a latent space. Operating on the latent space, M number of independent decoders generate M output images.

For the specific application of I^n2I, two generators are used for the forward and reverse transformations. When there are n input images, M is set to be equal to one during the forward transformation where the goal is to generate a single output image ($N = n, M = 1$). In the reverse transformation, a single input image is processed to generate n outputs thereby making $N = 1$ and $M = n$. Therefore, generator networks used in I^n2I are asymmetric in structure as shown in Figure 1 (a) (bottom).

The proposed method treats n inputs independently initially in the forward transformation and then extracted features are fused together. The fused feature is first transformed into a latent space Z as shown in in Figure 1 (a) (bottom) and then transformed into the target domain. In the reverse transformation, the single input is mapped back to the same latent space first. Then, the latent space representation is used to produce n outputs belonging to n source domains. In this formulation, $n + 1$ discriminators are used, one for each domain as opposed to CycleGAN. In addition, a latent space consistency loss is added to ensure that the same concept in all domains have a common latent space representation.

Problem Formulation. Formally, given n number of input modalities $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$, the objective is to learn a transformation $f_{\mathcal{S} \to T}(.)$. Here, we note that the input to the forward transformation is a set of images, where the output of the transformation is a single image. Similarly, the backward transformation $f_{T \to \mathcal{S}}(.)$ takes a single image input and produces n output images.

In order to approach the solution to this problem, first we view all input images and the desired output image as different representations of the same concept. Motivated by the techniques used in domain adaptation [1],[5],[23] we hypothesize the existence of a latent representation that can be derived using the provided representations. With this assumption, we treat our original problem as a series of sub-problems where the requirement is to learn the transformation and the inverse transformation to the latent representation from each domain. If the latent representation is Z, we will attempt to learn transformations $h_{I \to Z}$ and $h_{Z \to I}$, where $I \in \{\mathcal{S}, T\}$ and $h_{I \to Z} = h^{-1}_{Z \to I}$. With this formulation, the forward transform $f_{\mathcal{S} \to T}$ becomes $f_{\mathcal{S} \to T}(.) = h_{Z \to T}(h_{\mathcal{S} \to Z}(.))$ and the reverse transformation $f_{T \to \mathcal{S}}$ becomes $f_{T \to \mathcal{S}(.)} = h_{Z \to \mathcal{S}}(h_{T \to Z}(.))$.



(a) (b)

Fig. 1. (a) Network structure used for unsupervised image-to-image translation. Top: CycleGAN, Bottom: Proposed method for $I^n 2I$. (b) Multi-modal generator: generalization of the generator for multiple inputs and multiple outputs.

Adversarial Loss. In order to learn transformation $f_{\mathcal{S}\to T}$, we use an adversarial generator-discriminator pair $\{f_{\mathcal{S}\to T}(.), D_T(.)\}$ [6]. Denoting the data distributions of domains $\mathcal{S}$ and T as $P_{data}(s)$ and $P_{data}(t)$, respectively, the generator function tries to learn the transformation $f_{\mathcal{S}\to T}$. The discriminator is trained to differentiate real images from the target domain $\mathcal{S}$ from generated images $f_{\mathcal{S}\to T}(s)$. This procedure is captured in the adversarial loss as follows:

$$L_{GAN,\mathcal{S}\to T} = \mathbb{E}_{t\sim p_{data}(t)}[\log D_T(t)] + \mathbb{E}_{s\sim p_{data}(s)}[1 - \log\ D_T(f_{\mathcal{S}\to T}(s))]. \quad (1)$$

Similarly, to learn $f_{T\to S}$ we use a single generator $f_{T\to S}$. However, since there exists n input domains in total, we require n discriminators $\{D_{s_i}\}$, where $i = 1, 2, \ldots, n$, one for each domain. With this formulation, the total adversarial loss in backward transformation becomes a summation of n adversarial terms as follows:

$$L_{GAN,T\to\mathcal{S}} = \sum_{i=1}^{n} \mathbb{E}_{s_i\sim p_{data}(s_i)}[\log D_{S_i}(s_i)] + \sum_{i=1}^{n} \mathbb{E}_{t\sim p_{data}(t)}[1 - \log D_{S_i}(f_{T\to S_i}(t))]. \quad (2)$$

Latent Consistency Loss. As briefly discussed above, the adversarial loss only ensures that the generated image looks realistic in the target domain. Therefore, adversarial loss alone is inadequate to result in a transformation which preserves semantic information of the input. However, based on the assumption that both input and target domains share a common latent representation, it is possible to enforce a more strict constraint to ensure semantics between the input and the output are preserved. This is done by forcing the latent representation obtained during the forward transformation to be equal to the latent representation obtained during the reverse transformation for the same input.

More specifically, for a given input s, a set of latent representations $h_{s\to Z}(s)$ are recorded. Then, this recorded vector is compared against the latent representation obtained during the reverse transformation $h_{T\to Z}(f_{\mathcal{S}\to T}(s))$. The latent consistency loss in the forward transformation is defined as,

$$L_{latent,\mathcal{S}\to T} = \mathbb{E}_{s\sim p_{data}(s)}\|h_{\mathcal{S}\to Z}(s) - h_{T\to Z}(f_{\mathcal{S}\to T}(s))\|_1. \quad (3)$$

Similarly, the latent consistency loss in the reverse transformation is defined as,

$$L_{latent,T\to\mathcal{S}} = \mathbb{E}_{t\sim p_{data}(t)}\|h_{T\to Z}(t) - h_{\mathcal{S}\to Z}(f_{T\to\mathcal{S}}(t))\|_1. \quad (4)$$

Cycle Consistency Loss. If the input and the transformed image do not share semantic information, it is impossible to regenerate the input using the transformed image. Therefore by forcing the learned transformation to have a valid inverse transform, it is further possible to force the generated image to share semantics with the input. Based on this rationale, in [30] cycle consistency loss is introduced to ensure that the transformed image shares semantics with the input image. Since this argument is equally valid for the multi-input case, we adopt cycle consistency loss [30] in our formulation. Proposed backward cycle consistency loss is similar to that of [30] in definition. We define the reverse cycle consistency loss as:

$$L_{cyc,T\to\mathcal{S}} = \mathbb{E}_{t\sim p_{data}(t)}[\|f_{\mathcal{S}\to T}(f_{T\to\mathcal{S}}(t)) - t\|_1]. \quad (5)$$

However, in comparison, the forward cycle consistency loss takes into account n inputs and compares the distance among the n reconstructions as opposed to [30]. The forward cycle consistency loss is defined as,

$$L_{cyc,s\to T} = \mathbb{E}_{s\sim p_{data}(s)}[\|F_{T\to s}(F_{s\to T}(s)) - s\|_1]. \quad (6)$$

Cumulative Loss. The final objective function is the addition of all three losses introduced in this section. The cumulative loss L_{total} is defined as follows:

$$L_{total} = L_{GAN,\mathcal{S}\to T} + L_{GAN,T\to\mathcal{S}} + \lambda_1(L_{cyc,T\to\mathcal{S}} + L_{cyc,\int\to T}) + \lambda_2(L_{latent,\mathcal{S}\to T} + L_{latent,T\to\mathcal{S}}), \quad (7)$$

where, λ_1 and λ_2 are constants.

Limiting Case. It is interesting to investigate the behavior of the proposed network in the limiting case when $n = 1$. In this case, both the number of input and output modalities of the network becomes one; i.e. $N = 1$ and $M = 1$. Therefore $\mathcal{S}$ becomes S in equations (1), (2), (5) and (6). In addition, with $n = 1$, summation in (2) reduces to a single statement. If we disregard the latent consistency loss by forcing $\lambda_2 = 0$, the

total objective reduces to,

$$\begin{aligned} L_{total} = & \mathbb{E}_{t\sim p_{data}(T)}[\log D_T(t)] + \mathbb{E}_{s\sim p_{data}(S)}[\log D_S(s)] \\ & + \mathbb{E}_{s\sim p_{data}(s)}[1 - \log \ D_T(f_{s\to T}(s))] \\ & + \mathbb{E}_{t\sim p_{data}(T)}[1 - \log D_S(t)] \\ & + \mathbb{E}_{t\sim p_{data}(t)}[\|f_{s\to T}(f_{T\to s}(t)) - t\|_1] \\ & + \mathbb{E}_{s\sim p_{data}(s)}[\|F_{T\to s}(F_{s\to T}(s)) - s\|_1]. \end{aligned}$$

This reduced objective is identical to the total objective in CycleGAN. Therefore, in the limiting case when $n = 1$, the proposed method reduces to the cycleGAN formulation when the latent consistency loss is disregarded.

Network Architecture. In this section, we describe the network architecture of the proposed Generator by considering the case where two input modalities are used; i,e when $n = 2$. The resulting two generators in this case is illustrated in Figure 2. It should be noted that the Convolutional Neural Network (CNN) architectures used in both forward and reverse transformations here are in coherence with the generic structure shown in Figure 1 (b). In principle, the generator can be based on any backbone architecture. In our work, we used ResNet [7] with nine resnet blocks as the backbone. In our proposed network, a CNN is used for each module in Figure 1(b). These CNNs are typically convolutions/transposed convolutions followed by nonlinearities, batch-normalization layers and possibly with skip connections.

Two input images (from the two input domains) are present as the input of the forward transformation. These images are subjected to two parallel CNNs to extract features from each modality. Then, the extracted features are fused to generate an intermediate feature representation. In our work, feature fusion is performed by concatenating feature maps of feature extraction stage and using a convolution operation to reduce the dimension. This feature is then subjected to a set of convolution operations to arrive at the latent space. Finally, the latent space representation is subjected to a series of CNNs with transposed convolution operations to generate a single output image (from the target domain).

During the backward transformation, a single input is present. A CNN with convolution operations is used to transform the input into the latent space. It should be noted that since there is only a single input, there is no notion of fusion in this case. Two parallel CNNs consisting of transposed convolutions branch out from the latent space to produce two outputs corresponding to domains $S1$ and $S2$.

This architecture can be extended n modalities. In this case, the core structure will be similar to that of Figure 2 except that there will be n parallel branches instead of two at either ends of the network. For the discriminator networks we use PatchGANs proposed in [8].

IV. Experimental Results

We test the proposed method on two publicly available multi-modal image datasets across two tasks against state-of-the-art unsupervised image-to-image translation methods. The training was carried out adhering to principles of unsupervised learning[1]. Even when ground truth images of the desired translation were available, they were not used during training. When available, ground truth images were used during testing to quantify the structural distortion introduced by each method through calculating PSNR and SSIM [26] metrics.

As the benchmark for performance comparison, we use CycleGAN [30] and UNIT [12] frameworks. Since both of these methods are specifically designed for single inputs, we used all available image modalities, one at a time to produce the corresponding outputs. In addition, we present the following two additional baseline comparisons: **1. CycleGAN (Concat).** Input of multiple modalities are concatenated as channels. Operating on the concatenated input, cycleGAN is used to find the relevant transformation. **2. CycleGAN (Wavelet).** Input images are first fused using a wavelet-based image fusion technique. Then, CycleGAN is operated on the fused image.

In the implementation of the proposed method, λ_1 and λ_2 in (7) are set equal to 10 and 1, respectively. Learning is performed using the Adam optimizer[10] with a batch size of 1. Initial learning rates of generators and discriminators were set equal to 0.0002 and 0.0001, respectively. Training was conducted for 200 epochs, where learning rate was linearly decayed in the last 100 epochs.

Image Colorization. The EPFL NIR-VIS dataset [2] includes 477 images in 9 categories captured in the RGB and the Near-infrared (NIR) image modalities across diverse scenes. Scenes included in this dataset are categorized as country, field, forest, indoor, mountain, old building, street, urban and water. We use this dataset to simulate the image colorization task. We generated greyscale images from the RGB visible images and use greyscale and NIR images as the input modalities with the aim of producing the corresponding RGB image. We randomly selected 50 images to be the test images and used the remaining images for training.

First we trained CycleGAN [30] and UNIT [12] models for each input modality independently. Then, the proposed method was used to train a model based on both input modalities. Obtained results for these cases are shown in Figure 3. Obtained PSNR and SSIM values for each method on the test data are tabulated in Table I. By inspection, CycleGAN operating on greyscale images and both images were able to identify segments in the image but failed to assign correct colors. For example, in the first row, the tree is correctly segmented but with a wrong color. In comparison, CycleGAN with NIR images have resulted in a much better colorization. Since the amount of energy a color reflects depends on the wavelength of the color, a NIR signal contains some information about the color of the object. This could be the reason why NIR images have performed better colorization compared to greyscale. The same trend can be observed in the outputs of the UNIT method.

On the other hand, the proposed method has produced a colorization very similar to the ground truth. As an example

[1]Code is available at https://github.com/PramuPerera/In2I

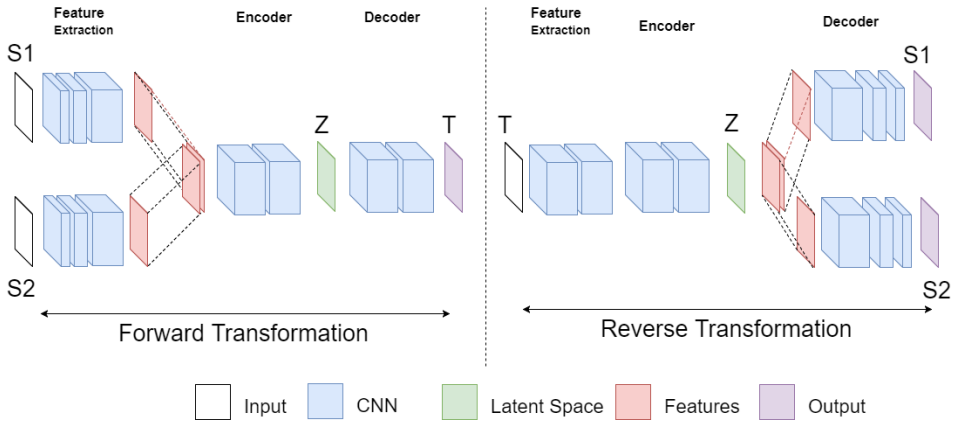


Fig. 2. Generator architecture of for $I^n 2I$ when two input modalities are used.

we wish to draw the attention of the reader to the color of the tree and the field in the first row, colors of the building and the tree in the last row. It has also recorded a superior PSNR and SSIM values compared with the other baselines as shown in Table I. It should be noted that PSNR and SSIM values only reflect how well the structure of objects in images have been preserved. It is not meant to be an indication of how well colorization task has been carried out.

TABLE I
QUALITATIVE EVALUATION OF COLORIZATION AND HYPERSPECTRAL-TO-REAL IMAGE TRANSLATION TASKS.

EPFL NIR-VIS		
Method	PSNR	SSIM
Ours (NIR+Grey)	**23.113 (9.147)**	**0.739 (0.008)**
UNIT (Grey)	8.324 (2.219)	0.041 (0.018)
UNIT (NIR)	15.331 (9.088)	0.544 (0.012)
CycleGAN (Grey)	8.438 (2.939)	0.056 (0.018)
CycleGAN (NIR)	17.381 (9.345)	0.657 (0.018)
CycleGAN (Concat)	8.597 (2.653)	0.054 (0.025)
CycleGAN (Wavelet)	8.597 (2.653)	0.054 (0.025)

Synthetic-to-Real Image Translation. In this subsection, we experiment on generating real images using synthetic images. For this purpose, we use two datasets, Synthia [19] and CityScapes [3], respectively as the source and the target domains. The Cityscapes dataset contains images taken across fifty urban cities during daytime. We use 1525 images from the validation set of the dataset to represent the target domain in the synthetic-to-real translation task. The Synthia dataset contains graphical simulations of an urban city. For our work, we only use the summer day light subset of the dataset which includes 901 images for training. The Synthia dataset provides RGB image intensities as well as the depth information of the scene. Hence, we use these as the two input modalities.

Results are shown in Figure 4. In this particular task, UNIT method has only changed the generic color scheme of the scene with incorrect association; for example note that skies look brown instead of blue in resulting images. In addition, objects in the scene continues to possess the characteristics of synthetic images. In contrast, CycleGAN has attempted to convert appearance of synthetic images to real. However, in the process it has distorted the structure of objects. When only depth information is used, the cycleGAN method is unable to preserve the structure of objects in the scene. For example, lines along the roads have ended up being warped in the learned representation in Figure 4. The CycleGAN model based on the visible images preserves the overall structure to an extent. However, vital details are either missing or misleading. For example, pavements are missing from images shown in rows 2 and 3 in Figure 4. The absence of a shadow on the road in row 2, addition of clutter in the left pavement in row 3 and disappearance of the telephone pole in row 4 are some of the notable incoherences. We note that despite of having more information, CycleGAN(concat) has produced a similar output of that of CycleGAN(Visible). Comparatively, fusion of both visible and depth information using proposed method has resulted in a more realistic translation. It should be noted that synthetic-to-real translation is a challenging problem in practice and when certain concepts were missing in either of source or target domains, the model found it difficult to learn such concepts. For example, training images from Cityscape did not have zebra crossings in any of the images. Therefore, the concept of zebra crossings is not learned well by the model as shown in row 1.

Impact of Multiple Inputs. Two experiments performed in this section are of different levels of difficulty. The colorization task is challenging due to the availability of diverse scenes. As a result, a single modality was not able to perform colorization satisfactorily. In this case, multi-image-to-image translation was able to induce a high improvement in terms of visual quality by using two informative input modalities. The second case, synthetic-to-real image translation, is more challenging. We note that the depth modality in this case is not very informative since it leads to image constructions of sub-standard quality. In comparison, the RGB synthetic image modality resulted in better translations. Using both modalities has improved the visual quality of the output. But this improvement was marginal as compared to the case of the colorization task. In summary, multiple modalities generally improve the visual quality of the output image; specially when the translation is more challenging. However, the amount of improvement introduced was dependent on the informativeness

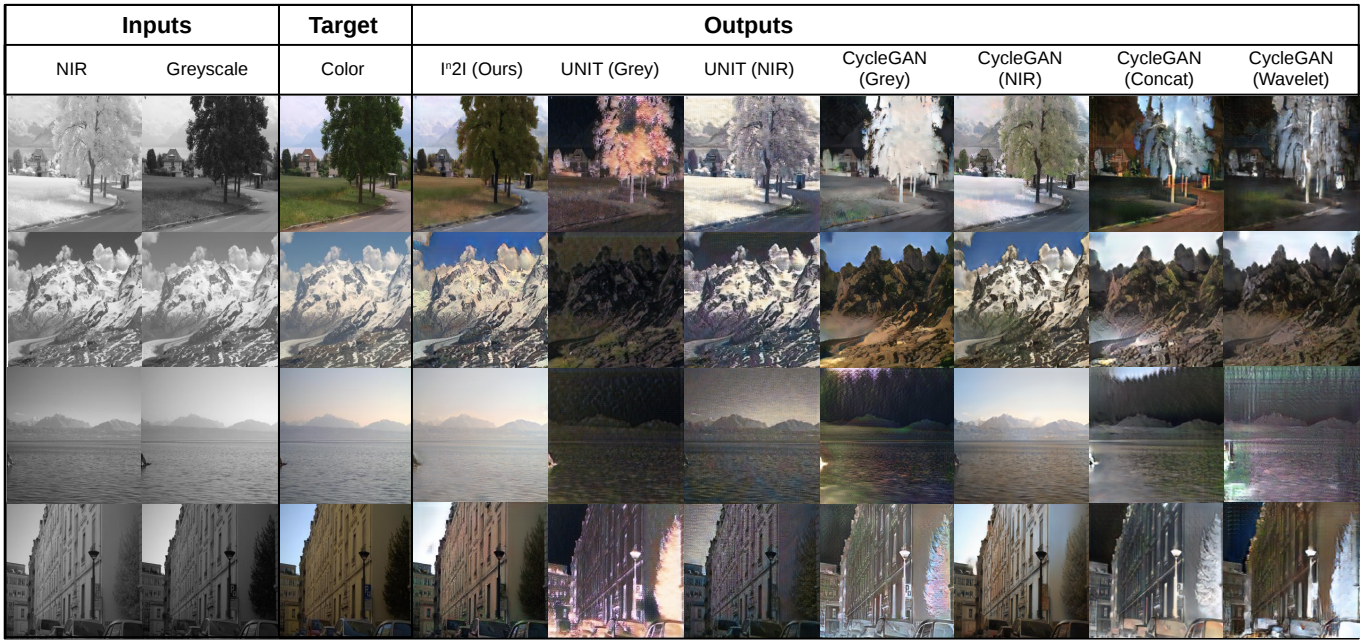


Fig. 3. Qualitative results for the image colorization task.

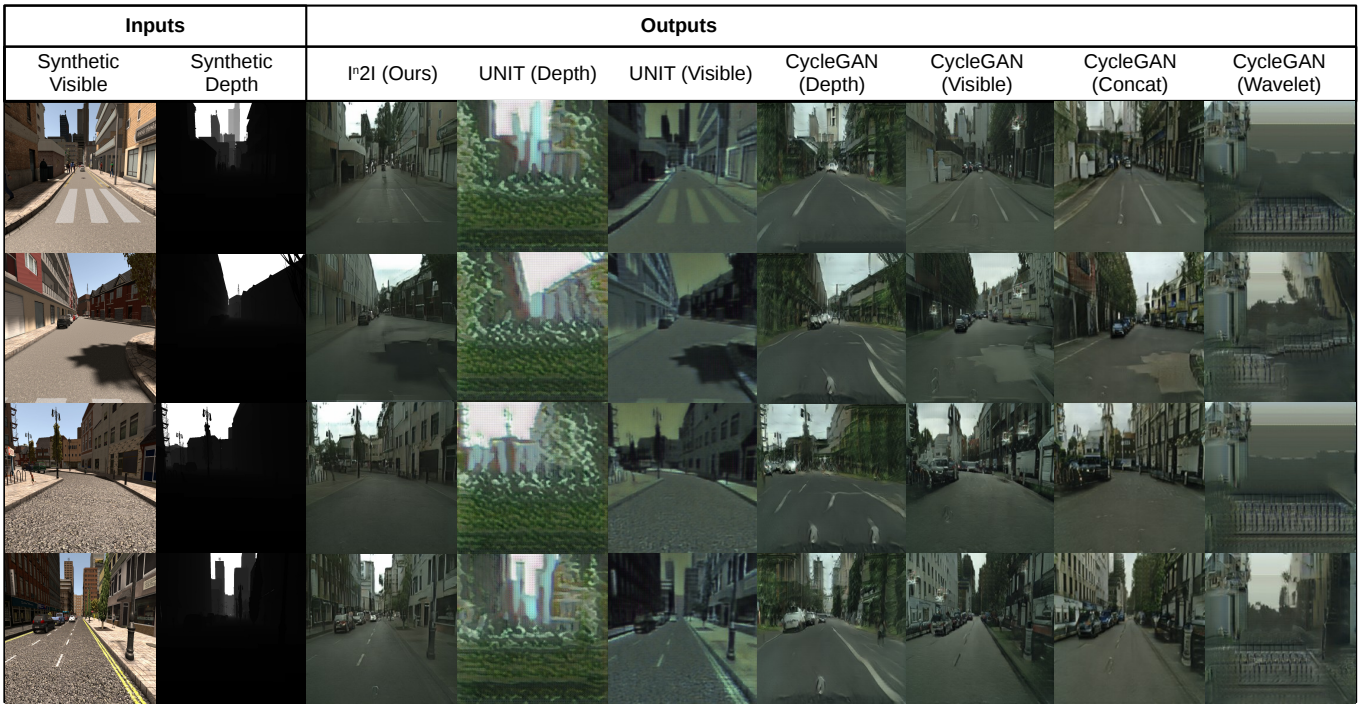


Fig. 4. Qualitative results for the synthetic-to-real translation task.

of the second modality.

Pixel-level Fusion vs Feature Fusion In the proposed network, information of input modalities are fused at the beginning of the encoder sub-network. In principle, fusion can carried out as pixel-level fusion, feature fusion or decision fusion [15]. Since the task in hand takes the form of image reconstruction, decision fusion is not applicable. In principle, it is also possible to use pixel-level fusion for this task as in CycleGAN (concat) and CycleGAN(Wavelet) methods. However, when the input modalities are from incompatible domains, pixel-level fusion results in incoherent reconstructions as shown in experimental results. In contrast, in the proposed method, images are first transformed into a latent space where both domains are compatible before fusion is performed. As a result the proposed method is able to produce images of higher perceptual quality compared with alternative fusion schemes.

V. Conclusion

We introduced multi-image-to-image translation problem. We proposed a multi-modal generator structure and a GAN

based framework as the initial direction to solve the problem. We tested the proposed method across two tasks against state-of-the-art unsupervised image-to-image translation methods. We showed that using multiple image modalities improves the visual quality of the output compared with results generated by the state-of-the-art methods. We analyzed the behavior of the proposed method in the limiting case and provided discussion as to when the use of multiple image modalities is most suited.

ACKNOWLEDGEMENT

This work was supported by US Office of Naval Research (ONR) Grant YIP N00014-16-1-3134.

REFERENCES

[1] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. Domain separation networks. 2016.

[2] M. Brown and S. Süsstrunk. Multispectral SIFT for scene category recognition. In *Computer Vision and Pattern Recognition (CVPR11)*, pages 177–184, Colorado Springs, June 2011.

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] X. Di and V. M. Patel. Face synthesis from visual attributes via sketch using conditional vaes and gans. *CoRR*, abs/1801.00077, 2018.

[5] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:59:1–59:35, 2016.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[8] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[9] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[10] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *The International Conference on Learning Representations*, 2015.

[11] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[12] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, 2017.

[13] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 469–477. Curran Associates, Inc., 2016.

[14] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[15] H. B. Mitchell. *Image Fusion: Theories, Techniques and Applications*. Springer Publishing Company, Incorporated, 1st edition, 2010.

[16] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, pages 689–696. Omnipress, 2011.

[17] G. Pajares and J. M. de la Cruz. A wavelet-based image fusion tutorial. *Elsevier Journal of Pattern Recognition*, 37:1855–1872, 2004.

[18] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[19] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[20] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems 29*, pages 2234–2242. 2016.

[21] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, 2017.

[22] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[23] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *AAAI Conference on Artificial Intelligence*, pages 2058–2065, 2016.

[24] L. Wang, V. A. Sindagi, and V. M. Patel. High-quality facial photo-sketch synthesis using multi-adversarial networks. *arXiv preprint arXiv:1710.10182*, 2017.

[25] X. Wang and A. Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, 2016.

[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *Transactions on Image Processing*, 13(4):600–612, Apr. 2004.

[27] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the International Conference on Computer Vision*, 2017.

[28] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.

[29] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *ECCV*, 2016.

[30] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *ICCV*, 2017.