IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 September 2023, accepted 28 September 2023, date of publication 5 October 2023, date of current version 11 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3322146

RESEARCH ARTICLE

Unsupervised Image to Image Translation With Additional Mask

HYUN-TAE CHOI[1], BONG-SOO SOHN[2], AND BYUNG-WOO HONG[1]
[1]Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, South Korea
[2]Department of Computer Science, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Byung-Woo Hong (hong@ai.cau.ac.kr)

This work was supported in part by the Institute for Information and Communication Technology Planning and Evaluation (IITP) grant funded by the Korean Government [Ministry of Science and ICT (MSIT)] through the Artificial Intelligence Graduate School Program, Chung-Ang University under Grant 2021-0-01341; in part by the National Research Foundation of Korea under Grant NRF-RS-2023-00251366; and in part by Chung-Ang University Research grant in 2023.

ABSTRACT With the development of deep learning, the performance of image-to-image translation is also increasing. However, most of the image-to-image translation models depend on the implicit method which does not explain why the models alter specific parts of the original input images. In this work, we assume that we can control the extent to which the models translate the input images using an explicit method. We explicitly create masks that will be added to the input images, aiming to highlight the difference between the inputs and the translated images. Since limiting the area of the masks directly affects the shape of the translated images, we can adjust the model through a simple regularization parameter. Our proposed method demonstrates that a simple regularization parameter, which regularizes the generated masks, can control where the model needs to change and remain. Furthermore, by adjusting the degree of the regularization parameter, we can generate diverse translated images from one original image.

INDEX TERMS Generative adversarial network, multi-modal image-to-image translation.

I. INTRODUCTION

The development of deep learning is highly related to the convolutional neural network [1], [2], especially in the image domain.The performance of the image-to-image translation is also enhanced with supervised-learning [4] and unsupervised-learning [5], [6], [7], [8], [9], [10]. In particular, due to the lack of the ground truth, unsupervised image-to-image translation has been actively studied with the development of generative adversarial network [14]. Recently, most works have solved this image-to-image translation using the generative adversarial network [5], [6], [7], [8], [9], [10]. However, the majority of them utilize the implicit method [5], [6], [7], [8], [9], [10] which cannot control the size of the area to be changed. To achieve the desired translated results that users want, the model needs to control the size of the regions by simply adjusting some regularization parameters. We address this problem by generating masks for each input image. For **explicit translation**, we add the generated masks with an auto-encoder [11]. Usually, auto-encoders consist of two parts: the encoder and decoder. When we optimize an auto-encoder with a specific loss function, we can obtain meaningful features from the encoder. Utilizing this concept, we generate masks that can be added pixel-wisely to the input images for image-to-image translation, following the equation: $\textit{Translated Output} = \textit{Mask} \oplus \textit{Input Image}$.

For the **control**, CycleGan [13] utilises a cycle-consistency loss to ensure that translated images retain the important features of the original input images. By employing this loss, we can impart certain characteristics of other domains to our generated masks. The generated masks are added to the input images and have values between -1 and 1. A pixel value of 0 in the mask implies that the mask does not alter the input image. We regularize the number of pixels with nonzero values in the mask. By controlling the number of nonzero values through a regularization parameter, we can produce diverse translated images(i.e., one-to-many mapping). In some cases, this enables us to obtain diverse backgrounds (e.g., if we do not restrict the regularization parameter, the background can change significantly). More details about the loss function are provided later in the paper.

The associate editor coordinating the review of this manuscript and approving it for publication was Nuno M Garcia.

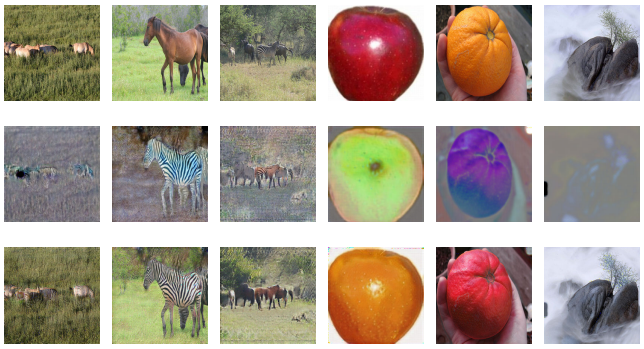


FIGURE 1. Some examples of our results about the horse2zebra, apple2orange and summer2winter dataset. The first row is the input, the second row is the generated mask and the last row is the translated output.

The contributions of this work are as follows:

- We address the multi-modal image-to-image translation problem using additional masks which can be intuitively employed.
- Our model allows controlling the size of the area that should be changed during translation.
- With our simple loss function, the model does not require a complex structure for the discriminator and generator.

In the remainder of this paper, we briefly present related works on image-to-image translation in Section II and then explain our network's structure with loss in Section III. Next, we show our image-to-image translation results in Section IV. Finally, we summarize and conclude our paper in Section V.

II. RELATED WORKS

A. GENERATIVE ADVERSARIAL NETWORK

Generative Adversarial Network (GAN) [14] has made significant contribution to unsupervised-learning, particularly in image generation problems. It consists of two parts: discriminator and generator. The discriminator and the generator are optimized in the concept of the zero-sum game problem. As the optimization progresses, the generator produces realistic images and the discriminator struggle to differentiate the real image and the generated image. GANs have also had a significant impact on image-to-image translation [5], [6], [7], [8], [9], [10] as well. In this case, the input images act as real images, and the translated image act as fake images for the discriminator. However, image-to-image translation faces challenges such as the hard optimization problem and mode-collapse problem, leading the model to generate the same translated image repeatedly(i.e. one-to-one mapping) [4], [5]. WGAN [15] used Wasserstein distance for optimization, and [16] designed Skip-Layer Excitation (SLE) instead of residual blocks of ResNet [3] for the deep neural network. Least Squares GAN(LSGAN) [18] calculated the $L-2$ norm between images and target labels to avoid the mode collapse and gradient vanishing, which can GAN training difficult. In our work, the loss function for the generation follows the scheme of the LSGAN. Usually, the discriminator of the GAN outputs a single value for the entire image, indicating whether it is real or fake. However, this single value might not capture high-frequency information in the generated images, as the discriminator is trained to learn the overall features of the target object in the input images. To ensure the generated image contain high-frequency information, PatchGan [4] was introduced. In PatchGan, the discriminator outputs a vector for one image, providing regional probability information. Additionally, PatchGan uses instance normalization [21] instead of the batch-normalization [22]. Our network also incorporates a PatchGan to capture high-frequency information in the generated images.

B. IMAGE-TO-IMAGE TRANSLATION

Image-to-Image translation aims to establish mapping functions between two different domains. Early studies addressed this task using supervised-learning. We show that our proposed method, which adds the generated masks to the input images, can be applied to the supervised learning scheme Fig. 2. We can get some proper solutions with Fig. 2. However, optimizing the models with the supervised learning requires a large number of ground truths, and even with ample dataset, their performance was limited due to the use of $L1$ loss and $L2$ loss between the ground truth images and the generated translated images, resulting in blurry results and a one-to-one mapping problem [4], [31], [32], [33], [34], [35]. Especially, Pix2Pix [4] argue that $L2$ loss leads to blurry results since $L2$ loss is minimized by averaging all plausible outputs. So [4] uses $L1$ loss instead of $L2$ loss. However, despite using the $L1$ loss, the blurriness issue still persisted [31], [32], [33], [34], [35]. To overcome these limitations, recent studies have explored unsupervised learning approaches. The classic solution is the variational autoencoder(VAE) [19], which learns the distribution of the input images. When we call the x as the input data and z as a latent, VAE learns $p(x|z)$ for the mapping function. VAE can be optimized with variational inference. This mapping function can be used for image-to-image translation. However, VAE's performance is not ideal as the distribution $p(x)$ is not directly obtained. GAN [14] can address this issue effectively. Pix2Pix [4] utilized $L1$ loss between the generated fake images and the real input images, but its supervised-manner led to blurry results for sparse input images. CycleGAN [13] can handle unpaired image datasets. There are two discriminators and two generators and it uses cycle-consistency loss to solve the mode collapse problem. Our work also used the cycle consistency loss for two unpaired image domains. InstaGAN [9] used additional instance information. It used the binary segmentation masks of the object in the input images as the instance information. It does not require cycle mapping between two different domains since it concatenates the segmentation masks to the input images. UNIT [5] assumes two different image domains can be mapped to a shared latent space, and it's loss consists of VAE loss and GAN loss. However, UNIT only can solve the one-to-one mapping. To apply the UNIT to the one-to-many

mapping problem, MUNIT [6] assumes that one image domain can be decomposed into a content space and a style space, and two image domains share a content space and each style spaces are domain-specific. References [10], [23], [24], and [25] used the concept of attention with GAN [24] to the image-to-image translation problem. Especially, AttentionGan [25] is solving image-to-image translation problems with a scheme very similar to ours. AttentionGan generates attention masks and content masks for the input image. They fuse the attention mask, content mask, and input to translate. Our work also generates a mask that is similar to the content masks of AttentionGan. But we do not use the attention masks, since attention masks highly limit the area that can be changed in the input image. Instead, we regularize the area with a simple loss. With this regularization for the areas of the generated masks, we can solve the one-to-many mapping. In summary, our proposed method adds generated masks to input images and can be applied to supervised learning and unsupervised learning as well. By employing a simple regularization loss for the generated masks, we achieve one-to-many mapping, allowing the model to control the size and extent of the translated regions effectively.

III. METHOD

A. NETWORK

In this section, we present two training schemes for our image-to-image translation method. To clarify, we first introduce the notations used in our approach:x represents an image in the first domain. y represents an image in the second domain. m_x, m_y are the generated masks for x and y, respectively. G_{yx}, G_{xy} refer to the generators responsible for generating m_x and m_y, respectively. Similarly, D_x, D_y are the discriminators used to discriminate images x and y, respectively. The first training scheme, as shown in Fig. 2, is trained in a supervised-manner. It involves one generator G which consists of an encoder and a decoder. The generator G generates a mask for the input image. We add the generated mask and the input image for translation: $\hat{y} = m_X \bigoplus x$. For Fig. 2, we only use L_1 loss between $\hat{y}$ and the ground truth y. With our simple additional mask, we can translate the input image to the other domain. However, this supervised scheme cannot handle one-to-many mapping and often results in blurry translated output. Moreover, this supervised approach requires a large number of ground truth images for effective optimization. To overcome this problem, we propose an unsupervised network. Our unsupervised method involves two generators and two discriminators, using the cycle-consistency loss for the optimization. Fig. 3 illustrates our proposed network structure(excluding the cycle-consistency part for simplicity) for translating the first image domain to the second image domain. The generator G_x consists of an encoder and a decoder. It generates a mask $m_x \in [-1, 1]$. To translate the input image $x \in X$ to the second domain image $\hat{y} \in Y$, we add x and m_x pixel-wisely: $\hat{y} = x + m_x$. To make the translated image $\hat{y}$ appear realistic, we treat $\hat{y}$ as a fake image, and y as a real image for the discriminator of the second domain. The translation from the second domain to the first domain follows a similar process as illustrated in Fig. 3. To ensure the generated masks effectively change the input images, the output range of the masks should be within $[-1, 1]$. If some pixels of the generated mask have a zero value, it implies that those parts of the input image do not need to be changed. We employ the structure from [13] for the encoder and the decoder, and we use the PatchGan discriminator to retain high-frequency information. For the patchGan discriminator, we used instance normalization [21].

B. LOSS FUNCTION

Equation (1) shows our loss for the optimization of the discriminator and the generator. We used the LSGAN scheme [18] between the discriminator's output and the target's label, for a stably good performance. Experimentally, the cross-entropy loss for the GAN was not working well. To overcome the aforementioned limitations of the supervised learning Fig. 2, we need to devise another loss function for the unsupervised scheme. For one-to-many mapping and to contain the original content, we used cycle-consistency loss [13] which is widely used in image-to-image translation problems. When we say the generated $\hat{y} = G_{YX}(x) + x$ and generated $\hat{x} = G_{XY}(y) + y$, our cycle-consistency loss is represented as (2).

$$\begin{aligned}\mathcal{L}_{lsgan} = &\mathbb{E}_{x\sim p_{data}(x)}(D_x(x)-1)^2 \\ &+ \mathbb{E}_{y\sim p_{data}(y)}(D_x(G_{XY}(y)+y))^2 \\ &+ \mathbb{E}_{x\sim p_{data}(x)}(D_y(y)-1)^2 \\ &+ \mathbb{E}_{y\sim p_{data}(y)}(D_y(G_{YX}(x)+x))^2 \qquad (1)\\ \mathcal{L}_{cycle} = &\mathbb{E}_{x\sim p_{data}(x)}\|G_{XY}(\hat{y})+\hat{y}-x\|_1 \\ &+ \mathbb{E}_{y\sim p_{data}(y)}\|G_{YX}(\hat{x})+\hat{x}-y\|_1 \qquad (2)\end{aligned}$$

But we did not use identity preserving loss (3) in [13] since our translated scheme consists of an additional process which is the sum of input and generated mask: $G_{YX}(x) + x$ and $G_{XY}(y) + y$. If we use identity preserving loss (3), the mask generator creates masks with all values consisting of zero which induces the work of not changing any part of the input images.

$$\begin{aligned}\mathcal{L}_{idt} = &\mathbb{E}_{x\sim p_{data}(x)}\|\|G_{YX}(x)+x-x\|\|_1 \\ &+ \mathbb{E}_{y\sim p_{data}(y)}\|\|G_{XY}(y)+y-y\|\|_1 \qquad (3)\end{aligned}$$

And to control the size of the region to be changed, we used the L_1 norm of the generated masks.

$$\begin{aligned}\mathcal{L}_{regularization} = &\mathbb{E}_{x\sim p_{data}(x)}\|\|G_{YX}(x)\|\|_1 \\ &+ \mathbb{E}_{y\sim p_{data}(y)}\|\|G_{XY}(y)\|\|_1 \qquad (4)\end{aligned}$$

Our total loss consists of (1), (2), and (4).

$$\mathcal{L}_{total} = \lambda_g \times \mathcal{L}_{lsgan} + \lambda_c \times \mathcal{L}_{cycle} + \lambda_r \times \mathcal{L}_{regularization} \qquad (5)$$

λ_c controls the cycle consistency and λ_r controls the degrees of the regularization. The shape of the masks highly correlated

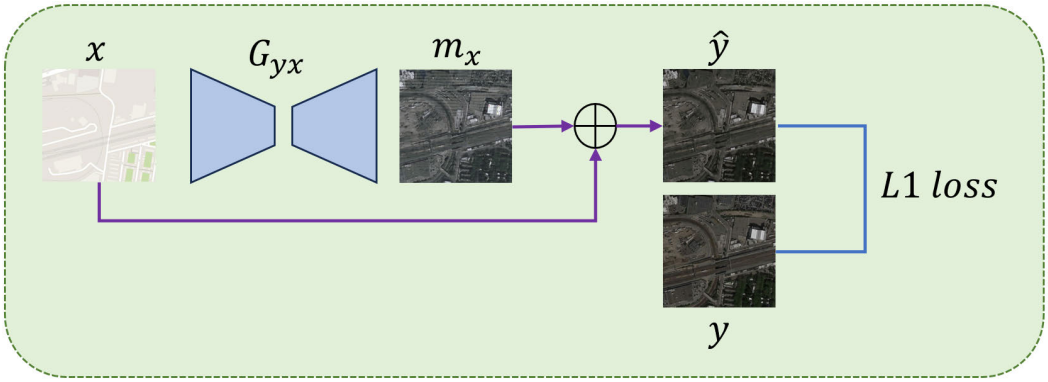


FIGURE 2. Illustration of our supervised network architecture. The Generator G consists of encoder and decoder. G generates a m that will be added to the input image x and it becomes $\hat{y}$ that is translated to the other domain. The loss for supervised network is calculated with L_1 loss between $\hat{y}$ and ground truth y.

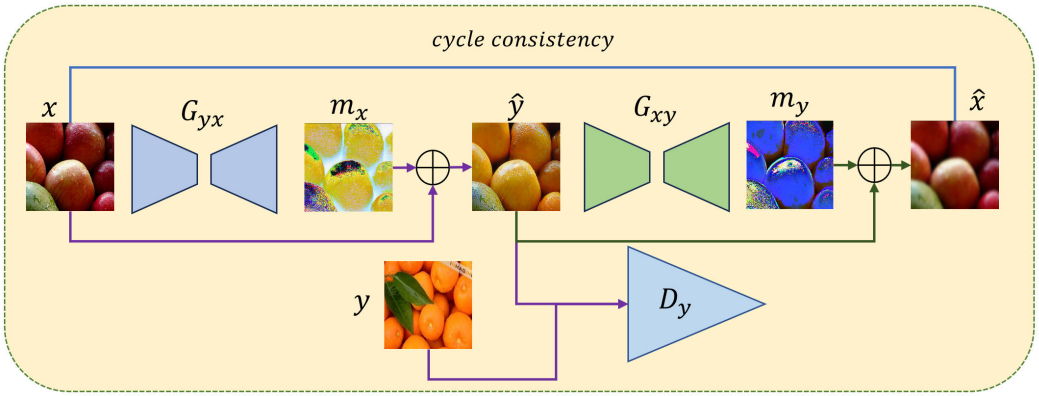


FIGURE 3. Illustration of our proposed network architecture for the first domain. The Generator G consists of encoder and decoder. G generates a m_x that will be added to the input image x and it becomes $\hat{y}$ that is translated to the second domain. D_y gives fake label to the $\hat{y}$ and gives real label to the second domain input image y.

to the λ_r. The larger λ_r, the smaller the size of the changing area which leads to a greater tendency to preserve the original. Conversely, when the λ_r is smaller, the degree of similarity between the translated image and the original diminishes. In the experiment section, we show the results according to the λ_r.

IV. EXPERIMENT

A. DATA

To show our method works well for the unsupervised image-to-image translation, we evaluate our method on some datasets: horse2zebra [13], maps [13], apple2orange [13], summer2winger [13], dog2cat and some synthetic circle and square images that we generate. For test, we split the train data into 8 to 2 ratios.

B. EXPERIMENTAL RESULTS

For the experiment, we used Adam optimizer [26] with a fixed learning rate of 0.0002 and the general value for $\beta_1 = 0.5$ and $\beta_2 = 0.999$. And all of our experiments are done with fixed values 0.5 for λ_g and 10 for λ_c in (5) since these two values have shown appropriate results. Similar to [13], random crop and left-right flip data augmentations are used. We also used the image replay buffer [13] with a probability of 50 percent.

1) ACCORDING TO THE REGULARIZATION PARAMETER

First, we compare the results of our method according to the degree of the regularization parameter. We control the λ_r in (5) to show the influence of our generated masks. In this experiment, we did not use the discriminator of PatchGan [4]. Instead, we rescaled the input image to 64×64 and used the same discriminator structure as [27]. This decision was made to focus solely on the main features that can differentiate two different image domains. It involves disregarding high-frequency information because reducing the image size achieves a similar effect to deblurring the image [36], [37]. The discriminator calculates whether the input images are real or fake based on the entire smaller images, using only low-frequency information about the objects. The batch size is 32. We evaluated our method using apple2orange [13], and a synthetic image dataset that we created for this experiment. Fig. 1 presents the

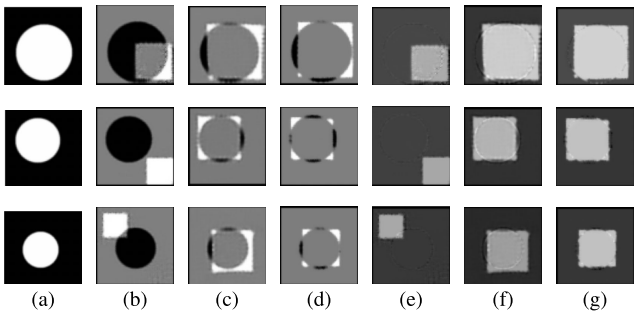


FIGURE 4. Image translation results of the synthetic images circle to square. (a) original circle image. (b) generated mask when $\lambda_c = 0.01$. (c) generated mask when $\lambda_c = 1$. (d) generated mask when $\lambda_c = 10$. (e) translated to square image when $\lambda_c = 0.01$. (f) translated to square image when $\lambda_c = 1$. (g) translated to square image when $\lambda_c = 10$.

results obtained based on the degree of the regularization parameter λ_r. To show the accurate role of the generated masks, we normalize the masks to $[0, 1]$ for the plot, considering that the originally generated masks are in $[-1, 1]$. The λ_r controls the area of the generated mask with the L_1 norm. As shown in Fig. 1, when the λ_r is small, large parts of the input image are translated, and the generated masks have many non-zero values, as seen in (b) and (e) of Fig. 1. On the contrary, when λ_r is large, the area of the changing part decreases, as evident in (d) and (g) of Fig. 1. The regions of the masks that aim to be changed are concentrated into four corners of the circle when optimized well with a sufficiently large value of x. For λ_r values slightly smaller than the optimal λ_r, the model still attempts to change around the four corner areas, but it also shows some deviation from the appropriate λ_r value, as observed in (c) and (f) of Fig. 1. However, if λ_r is too large, the generator generates unintended masks. Fig. 5 illustrates such unintended results when λ_r is set to 200. The translated output becomes square, but unlike (d) in Fig. 1, the generated mask is not concentrated on the corners of the circle. Note that there should be no edges of the circles in (e),(f), and (g) of Fig. 1 if the model makes perfect masks. However, as we mentioned above we rescale the input synthetic images to 64×64 to only change the overall shape, since in these synthetic images the low-frequency feature means the shape.

Fig. 6 shows intuitive results according to the λ_r on the apple2orange and horse2zebra datasets. As same as

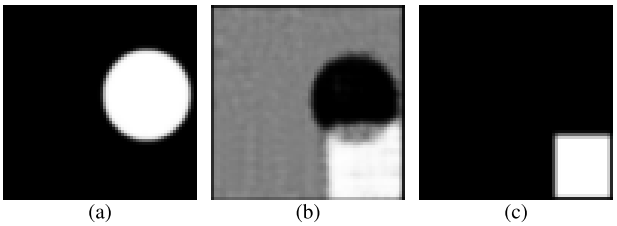


FIGURE 5. Image translation results of the synthetic images circle to square when λ_r is too large. (a) original circle image. (b) generated mask. (c) translated to square image.

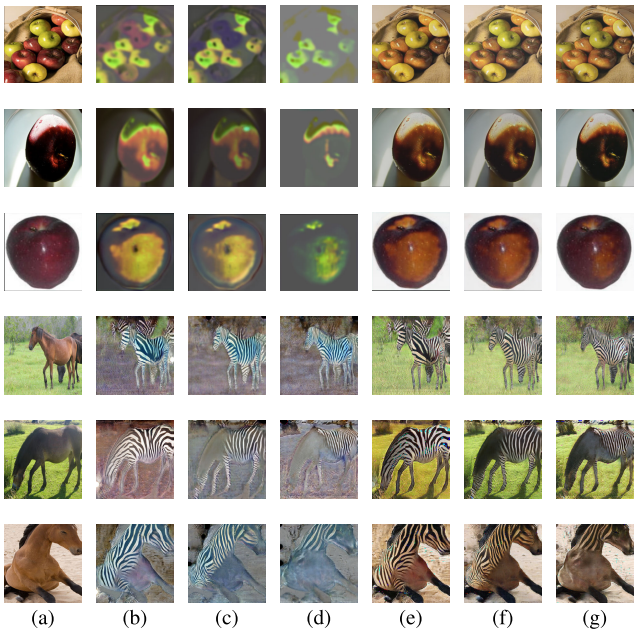


FIGURE 6. Image translation results of the apple2orange and horse2zebra. The three rows above are the results of the apple2orange and the three rows below are the results of the horse2zebra. (a) original input images. (b) generated masks when $\lambda_c = 0.01$. (c) generated masks when $\lambda_c = 0.1$. (d) generated masks when $\lambda_c = 1$. (e) translated images when $\lambda_c = 0.01$. (f) translated images when $\lambda_c = 0.1$. (g) translated images when $\lambda_c = 1$.

synthetic dataset, we rescale the apple2orange dataset to 64×64. However, since we cannot get proper results when using horse2zebra dataset, we use the high resolution of the horse2zebra image by rescaling to 256×256 and we used structures of [13] for the generator and [4] for the discriminator. The top three rows depict the results for the apple2orange, and the bottom three rows depict the results for the horse2zebra. In the apple2orange dataset, the main distinguishing feature between oranges and apples is their color, rather than the shape and texture. Because of this color feature, the λ_r controls the size of the regions that will be changed from red to yellow. The generated masks (b), (c), and (d) are the results of λ_r 0.01, 0.1, and 1, respectively. As λ_r increases, smaller regions in (a) are changed from red to yellow, and the translated outputs (g) have much smaller yellow regions than (e) and (f). The biggest difference between horses and zebras is the presence of stripes. With the results of the horse2zebra, we can see that as the λ_r increases, the number of stripes of the translated zebra increases. With Fig. 1 and Fig. 6, we can see that the simple L_1 loss for the regularization and the regularization parameter λ_r can control the regions of the input images to be changed. And if we assign proper value to λ_r, we can get reasonable masks and translated outputs.

2) OTHER RESULTS

In this part, we show the results of horse2zebra, summer2winter, and dog2cat. For this experiment, we used structures of [13] for the generator and [4] for the discriminator. To get the high-frequency information which means the details

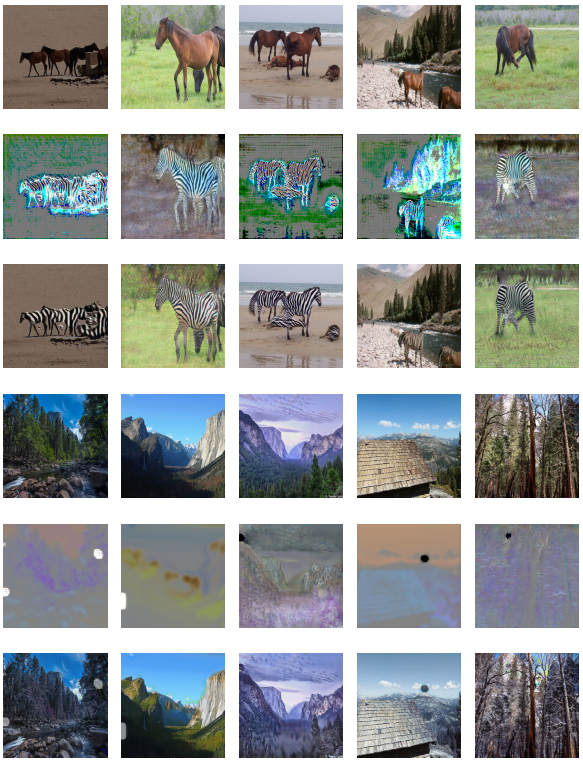


FIGURE 7. Image translation results of horse2zebra and summer2winter. The three rows above are the results about horse2zebra and the three rows below are the results about summer2winter dataset. The first row and fourth row are input images. The second row and the fifth row are the generated masks. And the third row and sixth row are the translated outputs.

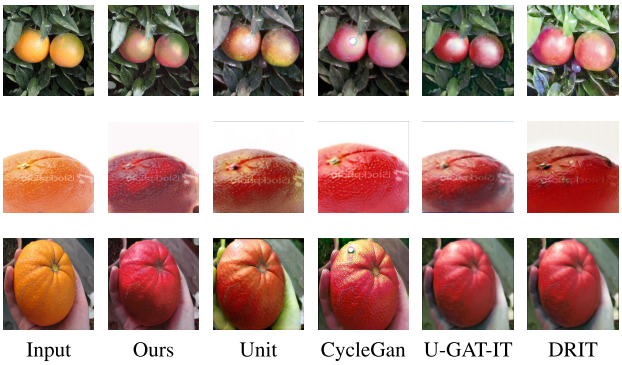


FIGURE 8. Different works for mapping orange to apple.

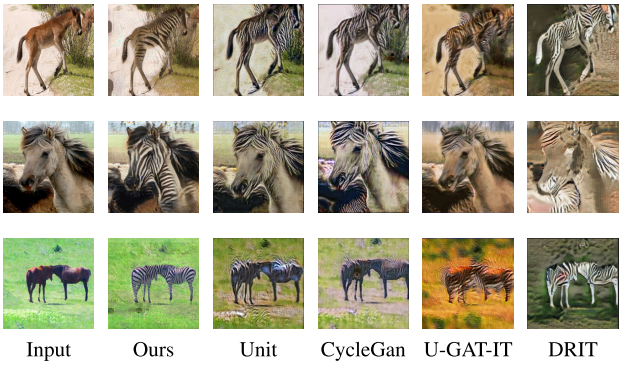


FIGURE 9. Different works for mapping horse to zebra.

of input images, we rescaled the input image to 256 × 256, and we tried many values of λ_r to get proper results. In Fig. 7, all the columns are the results with different λ_r values.

Fig. 7 shows the results of the horse2zebra and summer2winter datasets. The three rows above are the results about horse2zebra and the three rows below are the results of the maps dataset. The first row and fourth row are input images. The second row and the fifth row are the generated masks. And the third row and sixth row are the translated outputs. The third row is the result of the summation of the first row and the second row. And the sixth row is the result of the summation of the fourth row and the fifth row. With the three rows above, the masks only change the foreground which is considered as horses when the λ_r is properly selected. With the three rows below, we also can get the proper translated maps solutions.

3) COMPARE WITH OTHER WORKS

We conducted a comparison of our work with other existing methods Unit [5], CycleGan [13], U-GAT-IT [29], and DRIT [30] in Fig. 8, Fig. 9, and Fig. 10. Fig. 8, Fig. 9, and Fig. 10 are the results about orange to apple, horse to zebra, and winter to summer respectively. In Fig. 8, we can observe that all methods perform well for the orange-to-apple translation, as the primary distinguishing feature between oranges and apples is their color. However, some methods tend to make errors when there are objects in the image other than oranges. They have an error that changes the color of the object. In Fig. 9, our method outperforms other approaches. Stripes, which are not present in horses and exist only in zebras, exist only in horse areas in our results, while in other works, stripes appear even in places other than horse areas or are expressed only in a very small part of horses. In addition, compared to our results with little background change, background changes occur very severely in the results of other works. Fig. 10 is the result of the winter-to-summer translation. Due to the composition of the summer-to-winter dataset, the biggest

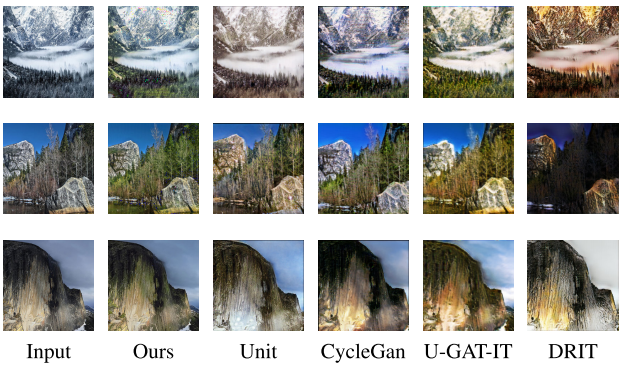


FIGURE 10. Different works for mapping winter to summer.

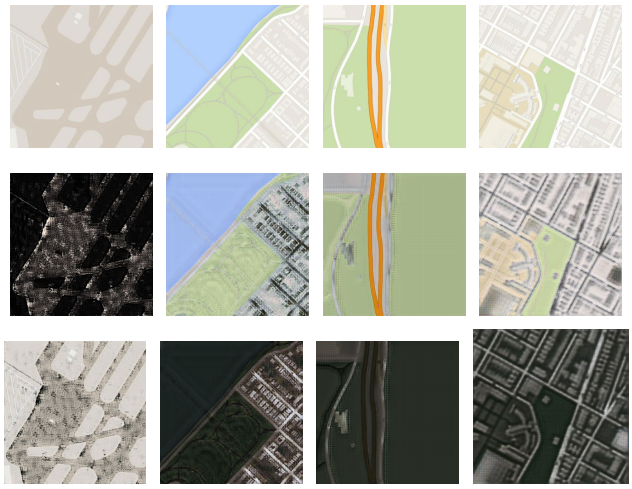


FIGURE 11. Failure results of maps dataset. The first row is the input, the second row is the generated mask, and the third row is the translated outputs.

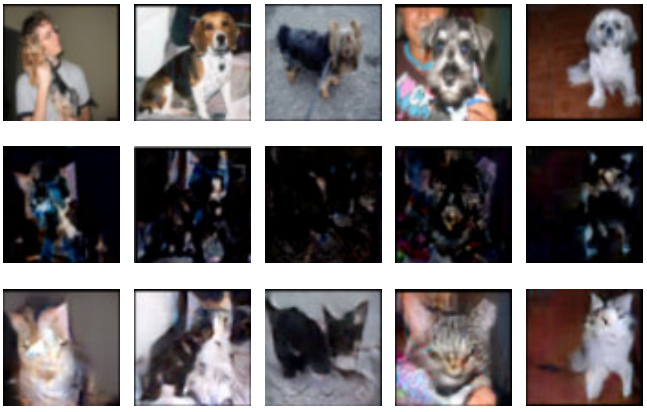


FIGURE 12. Failure results of dog2cat. The first row is the input, the second row is the generated mask, and the third row is the translated outputs.

difference between winter and summer is color (white for winter and green for summer). Our works are well-performing color translations from white to green. However as we can see with the other works' results, there are inappropriate changes in colors such as purple, pink, and blue. In addition, blurry effects could be observed from other works, and it can be seen that the phenomenon is more prominent in Unit and CycleGan.

4) FAILURE CASE

Fig. 7 shows the successful results with our proposed additional mask scheme. But we encountered many failure cases. Fig. 11 and Fig. 12 show the failure case when we try to translate the simple maps to complicate maps and try to translate dog to cat respectively. With Fig. 11 and Fig. 12, the additional mask scheme is not working well on complex datasets since this scheme is too explicit. When we see Fig. 11, the masks cannot make a detail of the buildings. They just generate the shape of the buildings. Also, Fig. 12 shows that the additional masks only change the shape of the dog: Lying

FIGURE 13. Image translation results of the dog2cat. (a) original dog image. (b) generated mask when $\lambda_r = 0.001$. (c) generated mask when $\lambda_r = 1$. (d) generated mask when $\lambda_r = 100$. (e) translated to cat image when $\lambda_r = 0.001$. (f) translated to cat image when $\lambda_r = 1$. (g) translated to cat image when $\lambda_r = 100$.

ears to sharp ears and make some stripes. The masks did not generate high-frequency information such as nose, eyes, and mouths.

Fig. 13 shows the results of the dog2cat dataset according to the value of λ_r. These results also fail to get some high-frequency information. However, we can notice that when we assign different $_r$, we can get different translated outputs. This means that we can manage one-to-many image-to-image translation problems with this simple regularization if we can solve the loss of high-frequency information.

V. CONCLUSION

In this paper, we propose a simple scheme which generates a mask that will be added to the input image for the image-to-image translation problem. This additional mask scheme can be trained in supervised-manner when we use L_1 loss between the translated output and the ground truth. Also, the proposed method can be trained in unsupervised-manner if we add cycle-consistency loss. Compared to previous works, our work produced similar or more reasonable results despite using a simpler network structure and simpler loss. In addition, unlike previous works that could not adjust the size of the area to be changed, our work can adjust it through the regularization parameter. What we haven't solved perfectly yet is that with simple data which have not too much high-frequency information, we can get properly translated output. On the other hand, if the data consists of high-frequency information, our additional scheme only acts to translate the shape features of the input image rather than texture and specific information. However, we can notice that if we control the regularization parameter, we can get various translated images for one input. Overcoming the limitation of our work could be valuable research in the future. Also, we expect that combining our scheme with the attention mechanism can adjust the specific area of the input image to be translated. If the research on this combination shows successful results, it will be able to effectively work not only on the general image-to-image translation but also on image-to-image translation through segmentation based on CNN.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
[2] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," 2015, *arXiv:1511.08458*.
[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[5] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 700–708.

[6] X. Huang, M. Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 172–189.

[7] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz, "Few-shot unsupervised image-to-image translation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10550–10559.

[8] K. Saito, K. Saenko, and M.-Y. Liu, "COCO-FUNIT: Few-shot unsupervised image translation with a content conditioned style encoder," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 382–398.

[9] S. Mo, M. Cho, and J. Shin, "InstaGAN: Instance-aware image-to-image translation," 2018, *arXiv:1812.10889*.

[10] Y. Lin, Y. Wang, Y. Li, Y. Gao, Z. Wang, and L. Khan, "Attention-based spatial guidance for image-to-image translation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 816–825.

[11] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proc. ICML Workshop Unsupervised Transf. Learn.*, 2012, pp. 37–49.

[12] M. J. Chong and D. Forsyth, "GANs N' roses: Stable, controllable, diverse image to image translation (works for videos too!)," 2021, *arXiv:2106.06561*.

[13] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[15] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[16] B. Liu, Y. Zhu, K. Song, and A. Elgammal, "Towards faster and stabilized GAN training for high-fidelity few-shot image synthesis," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–13.

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

[18] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2813–2821.

[19] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[21] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[23] X. Chen, C. Xu, X. Yang, and D. Tao, "Attention-GAN for object transfiguration in wild images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 164–180.

[24] D. Kastaniotis, I. Ntinou, D. Tsourounis, G. Economou, and S. Fotopoulos, "Attention-aware generative adversarial networks (ATA-GANs)," in *Proc. IEEE 13th Image, Video, Multidimensional Signal Process. Workshop (IVMSP)*, Jun. 2018, pp. 1–5.

[25] H. Tang, H. Liu, D. Xu, P. H. S. Torr, and N. Sebe, "AttentionGAN: Unpaired image-to-image translation using attention-guided generative adversarial networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 4, pp. 1972–1987, Apr. 2023.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[27] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[28] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 35–51.

[29] J. Kim, M. Kim, H. Kang, and K. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," 2019, *arXiv:1907.10830*.

[30] H. Y. Lee, H. Y. Tseng, J. B. Huang, M. Singh, and M. H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 770–785.

[31] C. Wang, H. Zheng, Z. Yu, Z. Zheng, Z. Gu, and B. Zheng, "Discriminative region proposal adversarial networks for high-quality image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 1530–1538.

[32] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

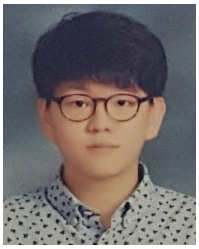
[33] A. Andonian, T. Park, B. Russell, P. Isola, J.-Y. Zhu, and R. Zhang, "Contrastive feature loss for image prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1934–1943.

[34] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. 14th Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 649–666

[35] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang, "MedGAN: Medical image translation using GANs," *Computerized Med. Imag. Graph.*, vol. 79, Jan. 2020, Art. no. 101684.

[36] N. van Noord and E. Postma, "Learning scale-variant and scale-invariant features for deep image classification," *Pattern Recognit.*, vol. 61, pp. 583–592, Jan. 2017.

[37] T. Lindeberg, *Scale-Space Theory in Computer Vision*, vol. 256. Cham, Switzerland: Springer, 2013.



HYUN-TAE CHOI received the B.S. and M.S. degrees in computer science from CAU. He is currently pursuing the Ph.D. degree in artificial intelligence. His current research interests include computer vision and machine learning.

BONG-SOO SOHN received the B.S. degree in computer science from Seoul National University, South Korea, and the M.S. and Ph.D. degrees in computer science from The University of Texas at Austin, in 2001 and 2005, respectively. He was a Postdoctoral Scholar with Prof. Taylor with Stanford University. He is currently an Associate Professor with the School of Computer Science and Engineering, Chung-Ang University, CAU. His research interests include multi-scale visualization, geometric modeling, and image processing from 3-D/4-D data with emphasis on bio-molecular and medical applications.

BYUNG-WOO HONG received the M.Sc. degree in computer vision from the Weizmann Institute of Science, in 2001, under the supervision of Prof. Shimon Ullman, and the D.Phil. degree in computer vision from the University of Oxford, in 2005, under the supervision of Prof. Michael Brady. He joined as a Faculty Member with the Computer Science Department, Chung-Ang University, CAU, in 2008, after his postdoctoral research with the Computer Science Department, University of California at Los Angeles, with Prof. Stefano Soatto. His research interests include image processing, computer vision, machine learning, and medical image analysis.