AMMUNIT: An Attention-Based Multimodal Multi-domain UNsupervised Image-to-Image Translation Framework

Lei Luo(✉) and William H. Hsu

Kansas State University, Manhattan, KS 66502, USA
{leiluoray,bhsu}@ksu.edu

Abstract. We address the open problem of unsupervised multimodal multi-domain image-to-image (I2I) translation using a generative adversarial network with attention mechanism. Previous works, such as CycleGAN, MUNIT, and StarGAN2 are able to translate images among multiple domains and generate diverse images, but they often introduce unwanted changes to the background. In this paper, we propose a simple yet effective attention-based framework for unsupervised I2I translation. Our framework not only translates solely objects of interests and leave the background unaltered, but also generates images for multiple domains simultaneously. Unlike recent studies on unsupervised I2I with attention mechanism that require ground truth for learning attention maps, our approach learns attention maps in an unsupervised manner. Extensive experiments show that our framework is superior than the state-of-the-art baselines.

Keywords: Image-to-image translation · Attention learning · Unsupervised learning

Image-to-image (I2I) translation refers to translating images from one domain to another featuring different styles, which are visually distinctive among different domains. An example is the task of turning images of cartoon sketches into real-life photographs. Many tasks in computer vision can be viewed as I2I translation, such as image inpainting [1], style transfer as in StyleGAN2 [2], and super-resolution [3]. Supervised I2I translation tasks need paired data sets that are costly to obtain, and such tasks are relatively easier to solve than their unsupervised counterpart. Under paired data supervision, I2I translation can be done by taking a regression approach [4] or using conditional generative models [5]. Our work addresses the more challenging unsupervised I2I translation task without access to paired data sets. Most of works on unsupervised I2I translation draw inspiration from CycleGAN [6] using the cycle consistency constraint, and have achieved impressive results. More recent studies, such as MUNIT [7] and StarGAN2 [8], have improved upon on CycleGAN and are able to translate images

Supported by Kansas State University.

E. Pimenidis et al. (Eds.): ICANN 2022, LNCS 13530, pp. 358–370, 2022.
https://doi.org/10.1007/978-3-031-15931-2_30

among multiple domains. These works, however, introduce unwanted changes to both objects of interest and the background, which is undesired. In our study we propose a simpler yet effective approach. Our framework only consists of one generator-discriminator pair and a mapping network, which enable multimodal and multi-domain translation. Moreover, our framework learns attention maps by using a attention module, which allows translating objects of interest and leave the background intact. Extensive experiments show that our framework is superior or comparable to state-of-the-art (SOTA) baselines. The contributions of our work can be summarized as follows:

- We propose a novel framework for unsupervised I2I translation with attention mechanism, which allows for image translation at instance level.
- Our framework learns attention maps with an unsupervised manner, which does not require segmentation annotations. Our attention module could be used as a plug-and-play add-on for existing pre-trained I2I translation frameworks, making them capable of learning attention maps at lower cost than training an attention module and its generator from scratch.
- Unlike previous works, such as MUNIT and DRIT [9], that require training $n(n-1)$ generators for translating images for n domains, we propose a novel framework architecture, which requires training only one generator-discriminator pair and achieves multimodal multi-domain I2I translation.
- Extensive experiments on publicly available data sets show that our framework is superior than SOTA baselines.

1 Related Work

Generative Adversarial Networks. Ideally, generative models learn how data is distributed, thus allowing data synthesis from the learned distribution. Since the advent of GANs [10], generative models have achieved impressive results in various tasks like image editing [11] and style transfer as in StyleGAN2. GANs try to learn the data distribution by approximating the similarity of distributions between the training data and the fake data produced by the learned model. GANs usually comprise a generator and a discriminator. The entire model learns by playing a minimax game: the generator tries to fool the discriminator by gradually generating realistic data samples, and the discriminator, in turn, tries to distinguish real samples from fake ones. GANs have been improved in various ways. To produce more realistic samples, an architecture of stacked GANs has been proposed: the laplacian pyramid of GANs [12]; layered, recursive GANs [13]; and style-based GANs (StyleGAN and StyleGAN2). Several studies have attempted to solve the instability training of GANs using energy-based GANs [14] and Wasserstein GANs [15]. In this study, we use GANs with their improved techniques to learn the distribution of data and how to translate among different domains.

Unsupervised I2I Translation. Unsupervised I2I translation translates images from one domain to another without paired data supervision. Much success in unsupervised I2I translation is due to the cycle consistency constraint, proposed in three earlier works: CycleGAN, DiscoGAN [16], and DualGAN [17]. To translate more than two domains, MUNIT and DRIT are proposed. These methods, however, sample style codes from a standard normal distribution, which leads to inferior translation results. Moreover, they require training $n(n-1)$ generators and n discriminators for translating images among n domains, which is computationally expensive and time-consuming. Our method proposes a simpler yet more effective approach that requires only one set of generator-discriminator. Recent systems such as StarGAN2 and ModularGAN [18] are developed to perform multimodal image-to-image translation to produce images with the same content but different contexts. All the aforementioned methods, however, introduce undesired changes to the background while translating images.

Attention Learning. Motivated by human attention mechanism, attention has been successfully applied in various computer vision and natural language processing tasks, such as machine translation [19], visual question answering [20], and image and video captioning [21]. Attention improves the performance of all these tasks by encouraging the model to focus on the most relevant parts of the input. In order to focus on the most discriminative semantic part and retain the background of images during translation, attention mechanism has been introduced into I2I. ConstrastGAN [22] takes a supervised approach and uses segmentation mask annotations as extra input data. Similar to our approach that learns attention masks without using extra annotation, AttentionGAN [23], ATAGAN [24], and AGGAN [25] add an attention module to each generator to locate the object of interest in image-to-image translation tasks. Thus, the background can be excluded from I2I translation. All these mentioned methods, however, are only able to translate two domains at a time. In order to remedy the drawbacks mentioned above, we propose an unified I2I translation framework with attention mechanism. Instead of having to train $n(n-1)$ generator-discriminator pairs for learning to translate among n domains, our methods only requires training one such pair. Thus, our framework reduces training time and memory footprint with better or comparable translation performance.

2 Methods

2.1 Preliminaries

Let x be an image that belongs to one of many domains. The diagram (a) in Fig. 1 shows an overview of our model. We start from a latent vector z that is sampled from a standard normal distribution. z goes through a mapping network, which learns style codes s of a specific domain, where m is a domain label and $s = M(z, m)$. Meanwhile, we employ a content encoder E_c to extract content codes c from image inputs. The decoder D takes content and style codes to generate reconstructed images x', which are then used by style encoder E_s to produce reconstructed style codes s'. We compute two L1 losses using the reconstructed images and style codes. Finally, we use a multi-task discriminator to distinguish real images from generated ones. During the translation phase, we keep the same content codes but use the style codes of target domains. Attention maps are learned using the attention module. Take translating a horse image x_m to a zebra image as an example, shown in the diagram (b) of Fig. 1. The horse image is processed by the encoder, resulting in style codes s_m and content codes c_m. In the meantime, the attention module extracts attention maps att from the horse image. The style codes of the zebra image s_n are exchanged with that of the horse image. Then, the decoder uses the content codes c_m and style codes s_n to generate an intermediate fake zebra image, whose background contains unwanted changes. We incorporate the attention map with the intermediate fake zebra image by $att \times D(c_m, s_n) + (1 - att) \times x_m$, which results in the final fake zebra image. Note that we only show the attention branch for translating horse to zebra due to space limitation, the other direction of translation is similar.

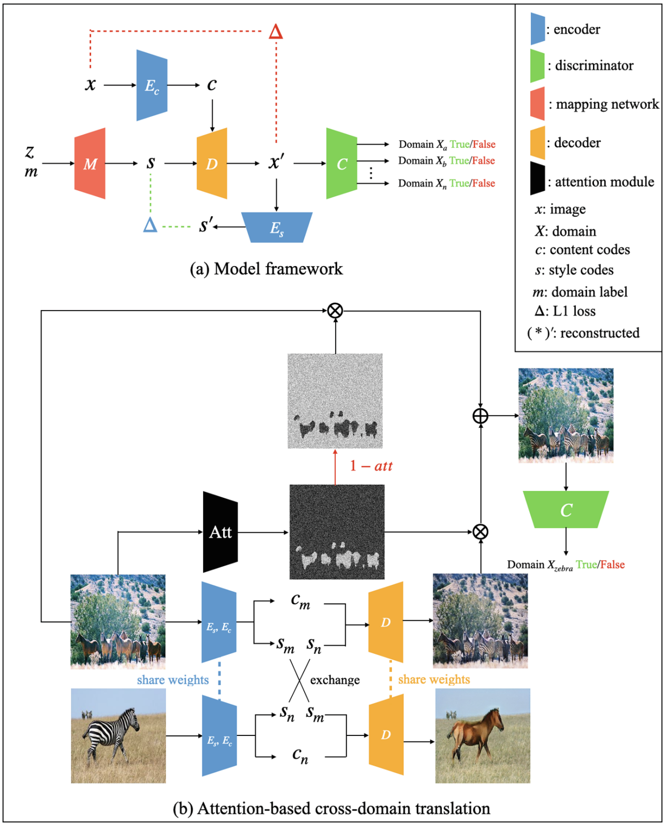


Fig. 1. The structure of our framework. (a) Shows how our framework learns, and (b) shows cross-domain translation within the horse and zebra domain. The attention branch of translating zebra2horse is similar to horse2zebra, and thus is not shown.

2.2 Framework Architecture

In this section, we outline the architecture of different modules in our framework.

Encoder. Our encoder has two sub-encoders: the style encoder and the content encoder. Both start with a convolution layer. The content encoder consists of six residual blocks [26]. All the layers are downsampled by average pooling operation (except for the last two layers) and are followed by an instance normalization (IN) [27]. The style encoder also comprises six residual blocks but without any activation function expect for the last residual block. Lastly, the style encoder consists of a convolution layer with leaky ReLU and a reshape operation before outputting style codes by the linear layer.

Mapping Network. Style codes of domains are modelled by a mapping network, which consists of eight linear layers with ReLU activation function expect for the last layer.

Decoder. The decoder maps latent codes, which consist of style codes and content codes, to the original image space. To apply style to images of different domain, the style codes are injected into the decoder by AdaIN [28] coupled with residual blocks. The last layer is a convolution layer whose outputs are generated images.

Attention Module. The attention module has an encoder-decoder architecture. The encoder consists three convolutional blocks, and the decoder has three convolutional layers with a sigmoid activation function at the end, which outputs the attention probability map.

Discriminator. The architecture of discriminator is similar to that of the style encoder except that it has one more convolutional layer to predict domains.

2.3 Training Objectives

In this section, we discuss the loss functions for learning our framework.

Image Reconstruction Loss. After images are encoded to style and content codes, the decoder maps the latent space back to the image space and reconstructs the image. Image reconstruction loss is formulated as:

$$L^{x}_{recon} = \|D(E_c(x), M(z, m)) - x\|_1 , \tag{1}$$

where m is the domain, to which image x belongs.

Style Code Reconstruction Loss. After encoding reconstructed images using the style encoder, we can obtain reconstructed style codes. We construct the style code reconstruction loss as follows:

$$L^{s}_{recon} = \|s - E_s(x')\|_1 , \tag{2}$$

where $x' = D(E_c(x), M(z, m))$ and $x \in X_m$.

Attention Consistency Loss. Images before and after translation should have the same attention maps. Thus, the attention consistency loss is defined as:

$$L_{att} = \left\| att(x_{mn}) - att(x_m) \right\|_1 , \tag{3}$$

where x_{mn} is the translated image, which is obtained by $att \times D(c_m, s_n) + (1 - att) \times x_m$. c_m is the content information of x_m and s_n is the style information of image x_n.

Regularization on Style and Content Codes. To further encourage style codes being domain-variant and content codes being domain-invariant, we add regularizers on style and content encoders. The style regularizer forces style codes of different domains to be different by minimizing L^s_{regu}, which is calculated as:

$$L^s_{regu} = - \left\| D(c_m, s_m) - D(c_m, s_n) \right\|_1 - \left\| D(c_n, s_m) - D(c_n, s_n) \right\|_1 , \tag{4}$$

where $(c_m, s_m) = (E_c(x_m), E_s(x_m))$ and $(c_n, s_n) = (E_c(x_n), E_s(x_n))$. c_m and s_m are content and style codes of image $x_m \in X_m$. c_n and s_n are content and style codes of image $x_n \in X_n$.

The content regularizer encourages content codes of different domains to be similar by minimizing L^c_{regu}, which is formulated as:

$$L^c_{regu} = \left\| D(c_m, s_m) - D(c_n, s_m) \right\|_1 + \left\| D(c_m, s_n) - D(c_n, s_n) \right\|_1 . \tag{5}$$

Inspired by StarGAN2, we calculate style diversity as:

$$L_{ds} = \left\| E_s(x_1) - E_s(x_2) \right\|_1 , \tag{6}$$

where $x_1 = D(E_c(x), M(z_1, m))$, and $x_2 = D(E_c(x), M(z_2, m))$, and z_1 and z_2 are two random latent vectors.

Adversarial Loss. GANs are used to match the distribution of translated results to real image samples, so the discriminator finds real and fake samples indistinguishable. We use two adversarial losses with one for learning latent-guided translation and the other for reference-guided translation. Latent-guided translation refers to using the mapping network to obtain target style codes, and reference-guided translation uses the style encoder to extract style codes of target domains. The adversarial loss for learning the discriminator C_m with latent-guided translation is formulated as:

$$\begin{aligned} L^l_{adv} = & \mathop{\mathbb{E}}_{z \sim N(0,I), x_n \sim p(X_n)} [log(1 - C_m(att \times D(c_n, M(z, m)) + (1 - att) \times x_n))] \\ & + \mathop{\mathbb{E}}_{x_m \sim p(X_m)} [log(C_m(x_m))], \end{aligned} \tag{7}$$

where m is the target domain label and the adversarial loss for learning the discriminator C_m with reference-guided translation is constructed as:

$$L^r_{adv} = \mathop{\mathbb{E}}_{x_m \sim p(X_m), x_n \sim p(X_n)} [log(1 - C_m(x_{nm}))] + \mathop{\mathbb{E}}_{x_m \sim p(X_m)} [log(C_m(x_m))], \tag{8}$$

where the discriminator C_m tries to tell if images are from the domain m, and x_{nm} is obtained by $att \times D(c_n, E_s(x_m)) + (1 - att) \times x_n$.

Full Objective. Our full objective is formulated as follows:

$$\begin{aligned} \underset{M,E,D}{min}\ \underset{C}{max}\ & \lambda_1 L^x_{recon} + \lambda_2 L^s_{recon} + \lambda_3 (L^s_{regu} + L^c_{regu}) \\ & + \lambda_4 (L^l_{adv} + L^r_{adv}) - \lambda_5 L_{ds} + \lambda_6 L_{att}, \end{aligned} \tag{9}$$

where λ_1 to λ_6 are hyperparameters for each loss term.

Model Training Scheme. We find it difficult for the model to converge when training the generator and the attention module simultaneously. Therefore, we first train the generator and the discriminator using $1e^{-4}$ as learning rate for 100,000 iterations, which is empirically calibrated. Then, we freeze the parameters of the generator when training the attention module for 30,000 iterations with the same learning rate. Lastly, we jointly train the entire framework for another 10,000 iterations using a smaller learning rate $5e^{-5}$.

3 Experiments

In this section we talk about the data sets, baselines, and evaluation metrics.

Baselines and Data Set. We compare our framework against four baseline models developed in recent years. CycleGAN is one of the pioneer work in unsupervised I2I, which is used as a baseline model. MUNIT and StarGAN2 achieve impressive results in unsupervised multimodal I2I translation, against which, thus, we compare our framework. For the sake of fair comparison, we compare our approach to AGGAN that is a recently proposed attention-based I2I translation framework.

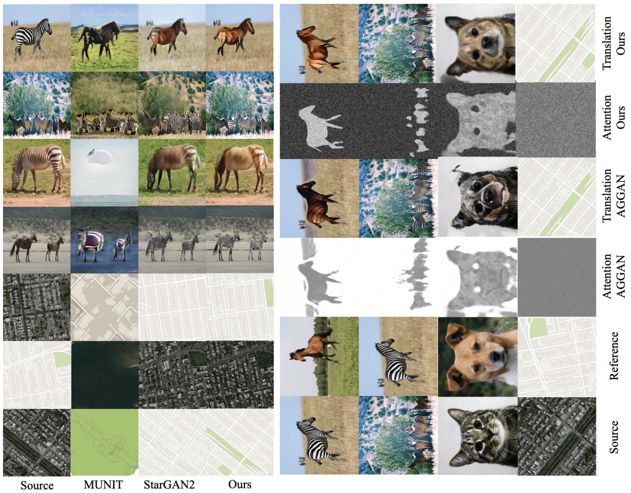
We evaluate our framework on the *horse2zebra*, *AFHQ*, and *map2aerial* data sets. The *horse2zebra* data set contains images of horses and zebras, and it is downloaded from ImageNet using keywords wild horse and zebra. There are in total 1,067 horse images and 1,334 zebra images are used for training, and 120 horse images and 140 zebra images are for testing. The *AFHQ* data set contains images of house cats, dogs, and wild animals (e.g. tigers, foxes, and lions). Similar to StarGAN2, we divide the *AFHQ* data set into domains of cats, dogs, and wild animals. The *map2aerial* data set are scraped from Google Maps, and images were sampled from in and around New York City. All images are of size 256 × 256.

Evaluation Metrics. We evaluate the visual quality of translation using the Amazon Mechanical Turk (AMT), which is based on user preferences given results of different models. To seek a quantitative measure that does not require human participation, Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) are employed similar to Chen et al. in AttentionGAN and AGGAN.

4 Results

In this section, we show the qualitative and quantitative results of the experiments. Ablation study is also carried out to evaluate the effectiveness of several key design choices.

Qualitative Results. We utilize the Amazon Mechanical Turk (AMT) to compare our results against the baselines based on user preferences. Given a source image and a reference image, we instruct AMT workers to select the best transfer result among all models. We ask 50 questions for all ten workers. As shown in Table 1, our method outperforms all the baseline models, especially for MUNIT, CycleGAN, and StarGAN2 that are not attention based I2I translation framework. Similar to MUNIT and StarGAN2, our model is also able to perform latent-guided and reference-guided translation. We illustrate examples of latent-guided translation in (a) of the Fig. 2, and Fig. 3 shows examples of I2I translation guided by reference images of all models. We can see that our model and AGGAN are capable of preserving the background information and only translating the objects of interests. CycleGAN and AGGAN are only able to perform reference-guided translation, thus their latent-guided translation results are not shown. We present two examples of attention maps of our model comparing against AGGAN in (b) of the Fig. 2, which shows that our attention maps are more accurate than AGGAN. From the results we argue that there should be a clear definition on what "undesired changes" are. It is clear that when performing translation, such as transferring a map into an aerial photo, we would assume the attention mask to be the entire image (See the attention map in figure (b) of the Fig. 2). We think it is probably more appropriate to apply such separation of background and background on domains of *horse2zebra* instead of *map2aerial*.



(a). Latent-guided translation

(b). Attention maps comparison

Fig. 2. (a) Are examples of latent-guided I2I translation results, and (b) compares attention maps generated by our framework and AGGAN.

Table 1. Votes from ATM workers for most preferred translation results.

Models	User preference (↑)
CycleGAN	8.31 %
MUNIT	2.55 %
StarGAN2	3.13 %
AGGAN	40.93 %
Ours	**45.08 %**

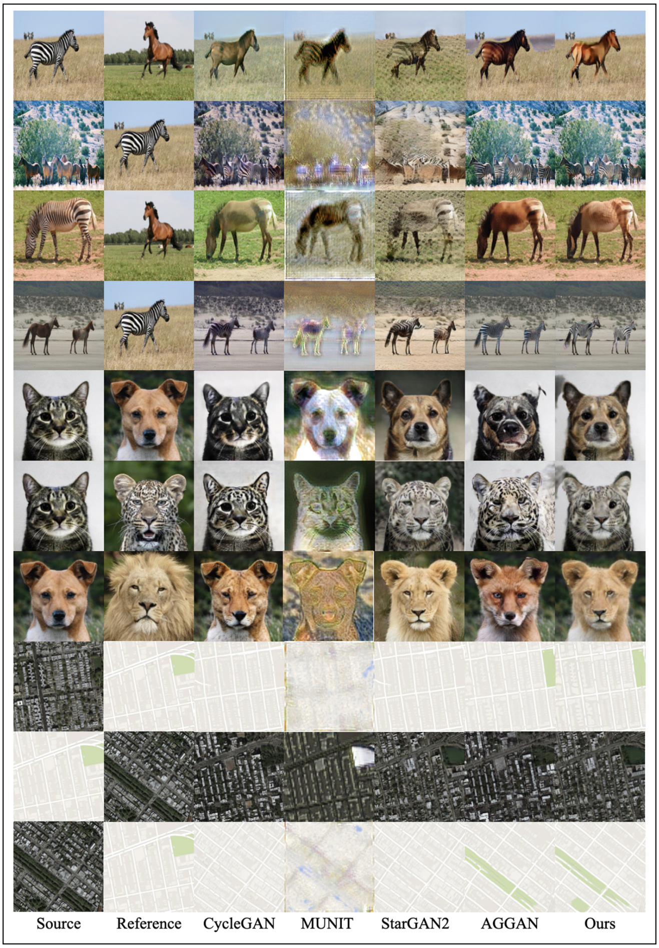


Fig. 3. Examples of reference-guided I2I translation by different models.

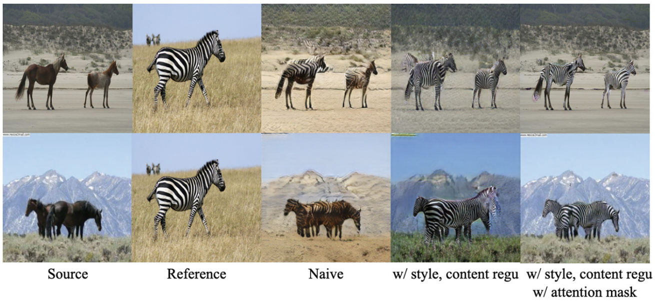


Fig. 4. An example of reference-guided translation by incrementally adding modules.

Quantitative Results. Similar to MUNIT and StarGAN2, our model is able to perform latent-guided and reference-guided translation. We evaluate all models using SSIM and PSNR, which require ground truth attention maps of images. Similar to AttentionGAN, we obtain attention maps using the DeepLab semantic image segmentation model [29] pretrained on MSCOCO [30] data set. Note that we only provide quantitative results on the *horse2zebra* data set because the DeepLab model is not trained on the *map2aerial* data set, and no ground truth attention maps are available for calculating SSIM and PSNR. As Table 2 and Table 3 show, our framework outperforms all baseline models, especially for CycleGAN, MUNIT, and StarGAN2 for a large margin. Again, CycleGAN and AGGAN are not capable of performing latent-guided translation. Therefore, quantitative results on these two models are not reported.

Table 2. Quantitative comparison on reference-guided translation.

Models	*horse2zebra*		*zebra2horse*	
	SSIM(↑)	PSNR (↑)	SSIM(↑)	PSNR (↑)
CycleGAN	0.7313	21.96	0.8453	26.31
MUNIT	0.1176	14.89	0.3664	15.29
StarGAN2	0.3281	16.86	0.4729	19.43
AGGAN	0.9686	33.16	0.9843	43.02
Ours	**0.9699**	**36.12**	**0.9851**	**44.11**

Table 3. Quantitative comparison on latent-guided translation.

Models	*horse2zebra*		*zebra2horse*	
	SSIM(↑)	PSNR (↑)	SSIM(↑)	PSNR (↑)
MUNIT	0.1925	11.66	0.3901	13.88
StarGAN2	0.3353	18.87	0.4953	19.92
Ours	**0.9712**	**33.76**	**0.9857**	**43.14**

Ablation Studies. To further validate effects of key design choices in our framework, we carry out ablation studies on the *horse2zebra* data set, whose results are shown in Table 4 and Fig. 4. Let the model without style, content regularizer, and attention module be the naive model. We can see that adding attention greatly helps increase translation results.

Table 4. SSIM and PSNR results of incrementally adding modules to our framework for reference-guided translation on the *horse2zebra* data set.

Modules	SSIM (↑)	PSNR (↑)
Naive model	0.3062	12.73
+ style, content regularizer	0.3511	19.04
+ attention masks	**0.9699**	**36.12**

5 Conclusions and Discussion

In this research, we present a simple yet effective attention-based framework for unsupervised I2I translation. Our framework not only translates solely objects of interests and leave the background unaltered, but also generates images for multiple domains simultaneously. Unlike similar studies on unsupervised I2I with attention mechanism that require ground truth for learning attention maps, our approach learns attention maps in an unsupervised manner. The qualitative and quantitative results show that our framework is superior than the SOTA baselines.

References

1. Wang, Y., Tao, X., Qi, X., Shen, X., Jia, J.: Image inpainting via generative multi-column convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 331–340. Curran Associates Inc., Montréal (2018)
2. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and Improving the Image Quality of StyleGAN. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, pp. 8110–8119 (2020)
3. Wang, Z., Chen, J., Hoi, S.C.H.: Deep learning for image super-resolution: a survey. IEEE Trans. Pattern Anal. Mach. Intell. 1 (2020)
4. Chen, Q-F., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1511–1520. IEEE, Honolulu (2017)
5. Isola, P., Zhu, J.-Y., Zhou, T.-H., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, pp. 5967–5976 (2017)
6. Zhu, J.-Y., Park, T., Isola, P., Efros A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251. IEEE, Venice (2017)

7. Huang, X., Liu, M.-Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Multimodal Unsupervised Image-to-Image Translation. LNCS, vol. 11207, pp. 179–196. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01219-9_11
8. Choi, Y., Uh, Y-J., Yoo, J., Ha, J-W.: StarGAN v2: diverse image synthesis for multiple domains. in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8185–8194. IEEE, Seattle (2020)
9. Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M., Yang, M.-H.: Diverse image-to-image translation via disentangled representations. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11205, pp. 36–52. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01246-5_3
10. Goodfellow, I. J., et al.: Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, pp. 2672–2680. MIT Press, Montreal (2014)
11. Zhu, J.-Y., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 597–613. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_36
12. Denton, E.L., Chintala, s., Szlam, A., Fergus, B.: Deep generative image models using a Laplacian pyramid of adversarial networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, pp. 1486–149. MIT Press, Montreal (2015)
13. Yang, J.-W., Kannan, A., Batra, D., Parikh, D.: LR-GAN: layered recursive generative adversarial networks for image generation. In: 5th International Conference on Learning Representations (ICLR), OpenReview.net, Toulon, France (2017)
14. Zhao, T., Mathieu, M., LeCun, Y.: Energy-based generative adversarial networks. In: 5th International Conference on Learning Representations (ICLR), OpenReview.net, Toulon, France (2017)
15. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML), pp. 214–223. PMLR, Stockholm (2017)
16. 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 (ICML), pp. 1857–1865. PMLR, Sydney (2017)
17. Yi, Z.-L., Zhang, H., Tan, P., Gong, M.-L.: DualGAN: unsupervised dual learning for image-to-image translation. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2868–2876. IEEE, Venice (2017)
18. Zhao, B., Chang, B., Jie, Z., Sigal, L.: Modular generative adversarial networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 157–173. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_10
19. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations (ICLR), OpenReview.net, San Diego, CA, USA (2015)
20. Yang, Z.-C., He, X.-D., Gao, J.-F., Deng, L., Smola, A.-J.: Stacked attention networks for image question answering. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21–29. IEEE, Las Vegas (2016)
21. Xu, K., et al.: Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning (ICML), Lille, France, pp. 2048–2057 (2015)

22. Liang, X., Zhang, H., Lin, L., Xing, E.: Generative semantic manipulation with mask-contrasting GAN. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11217, pp. 574–590. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_34
23. Chen, X., Xu, C., Yang, X., Tao, D.: Attention-GAN for object transfiguration in wild images. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11206, pp. 167–184. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01216-8_11
24. Kastaniotis, D., Ntinou, I., Tsourounis, D., Economou, G., Fotopoulos, S.: Attention-aware generative adversarial networks (ATA-GANs). In 13th IEEE Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), Aristi Village, Zagorochoria, Greece, pp. 1–5 (2018)
25. Tang, H., Xu, D., Sebe, N., Yan, Y.: Attention-guided generative adversarial networks for unsupervised image-to-image translation. In: International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, pp. 1–8 (2019)
26. He, K.-M., Zhang, X.-Y., Ren, S.-Q., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE, Las Vegas (2016)
27. Ulyanov, D., Vedaldi, A., Lempitsky, V.S.: Improved texture networks: maximizing quality and diversity in feed-forward stylization and texture synthesis. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4105–4113. IEEE, Honolulu (2017)
28. Huang, X., Belongie, S.-J.: Arbitrary style transfer in real-time with adaptive instance normalization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1510–1519. IEEE, Venice (2017)
29. Chen, L.-C., Papandreou, P., Kokkinos, I., Murphy, K., Yuille, A.-L: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Pattern Anal. Mach. Intell. 834–848 (2018)
30. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48