Avoiding Shortcuts in Unpaired Image-to-Image Translation

Tomaso Fontanini(✉), Filippo Botti, Massimo Bertozzi, and Andrea Prati

IMP Lab, Department of Engineering and Architecture, University of Parma, Parma, Italy
{tomaso.fontanini,massimo.bertozzi,andrea.prati}@unipr.it, filippo.botti2@studenti.unipr.it

Abstract. Image-to-image translation is a very popular task in deep learning. In particular, one of the most effective and popular approach to solve it, when a paired dataset of examples is not available, is to use a cycle consistency loss. This means forcing an inverse mapping in order to reverse the output of the network back to the source domain and reduce the space of all the possible mappings. Nevertheless, the network could learn to take shortcuts and softly apply the target domain in order to make the reverse translation easier therefore producing unsatisfactory results. For this reason, in this paper an additional constraint is introduced during the training phase of an unpaired image-to-image translation network; this forces the model to have the same attention both when applying the target domains and when reversing the translation. This approach has been tested on different datasets showing a consistent improvement over the generated results.

Keywords: Generative adversarial network · Attention generation

1 Introduction

Unpaired image-to-image translation aims at finding a mapping from an input image belonging to a source domain to an output image belonging to a target domain when a paired dataset of samples is not available. The preferred architecture for approaching this task is a generative adversarial network (GAN), where the generator is trained to apply the target domain to an input image and the discriminator is trained to distinguish if an image is real or generated. However, if no additional constraint is imposed during training, the generator could not only simply apply the target domain but also alter the overall shape/identity of the input in order to more easily fool the discriminator.

For this reason, CycleGAN [27] introduced a cycle consistency loss by adding an additional generator trained to learn an inverse mapping from the target domain back to the source domain. This solution allows to apply the target domains to an image without changing its overall content. Nevertheless, a

S. Sclaroff et al. (Eds.): ICIAP 2022, LNCS 13231, pp. 463–475, 2022.
https://doi.org/10.1007/978-3-031-06427-2_39

significant drawback introduced by the cycle consistency loss is that the generator could take a shortcut applying the target domain just enough to fool the discriminator, but also just as softly to make the reverse translation easier. A practical demonstration of this can be seen in Fig. 1, where the resulting images (second column) show a negligible transformation and therefore retain lots of information from the original domain (first column). This effect eases the reverse translation, but does not represent an *optimal* result, i.e. a result image which is indistinguishable from an image belonging to the target domain.

Fig. 1. Some samples that demonstrate the shortcuts introduced by the cycle consistency loss in CycleGAN for horse → zebra, orange → apple and apple → orange.

To solve this limitation, inspiration has been drawn by the recent applications of attention as valuable information to be used during the training of CNNs. The concept of attention in CNNs was first introduced by Zeiler and Fergus [23] as a way to visualize regions in the images that are important for the network when taking a certain decision or performing a certain task. Recently, attention was not only used as a mean of visual explanation of CNNs, but also actively during training. For example, attention was transferred from a teacher to a student model in order to improve the classification performance of the student in [22], attention maps were used effectively for semantic segmentation in [15] and, lastly, Dhar *et al.* [5] introduced an attention distillation loss for incremental learning that allowed to preserve the information about base classes when adding new ones, without storing any of their data. In addition to that, Liu *et al.* [16] showed that attention map can be generated effectively even in generative models like Variational AutoEncoders (VAE).

In this paper, we propose to actively use attention maps during the training of a CycleGAN. In particular, the intuition is that the attention obtained when translating an input image to a target domain and the attention obtained when translating the output image back to the source domain should be the same, because the network needs to focus on the same area of the network with the same intensity in both cases. This allowed to prevent the generator from taking shortcuts when applying the target domain, then resulting in images with a much higher quality.

To sum up, the main contributions of this paper are the following:

– A system that utilizes attention maps during the training of an unpaired image-to-image translation network allowing to limit the introduction of shortcuts caused by the cycle consistency loss. This improves the generated results without the need of any additional module;

– A quantitative and qualitative evaluation over common unpaired image-to-image translation datasets.

2 Related Work

Conditional GANs for Image-to-image Translation. Generative Adversarial Networks (GANs) were first introduced by Goodfellow *et al.* in [7] and since then they become very common in lots of deep learning applications. In addition to that, conditional GANs (cGANs) [19] allowed to achieve control over the generated samples by feeding the GAN model with additional information like labels [2] or text [24]. When both the input and the output of the generator are images, this is often referred to as *cGANs for image-to-image translation*. Initially, a great success was achieved using paired datasets of images [11], but very often it is not possible to have a ground truth when applying a target domain to an image.

For this reason, DiscoGAN [12] and, more notably, CycleGAN [27] were introduced. CycleGAN, which will be described in detail later in this paper, works by simultaneously training two generators and two discriminators. One generator is trained to produce images belonging to a domain Y starting from a domain X, while the other generator is trained in the opposite way. Also, CycleGAN uses a cycle consistency loss to force the output of a generator to be reversed back to its original domain when fed to the other generator, allowing to maintain the shape of the original input intact during the translation. Since then, the idea of introducing a cycle consistency loss for unpaired image-to-image translation has become very popular [4,9,10,14]. In addition to that, recently other authors noted that cycle loss can limit the efficacy of the translation task. Particularly, [20] relies on a council of networks that collaborates between each other and [25] proposes an adversarial-consistency loss for image-to-image translation. Yet, in our paper, cycle consistency loss is maintained and results are improved without the need of designing a completely-new architecture.

Attention Maps Generation. A very active research field consists in understanding how neural networks perform their tasks or take their decisions. Some preliminary results in this direction were achieved by [17] and [23]. After that, CAM (Combined Attention Model) was introduced by Zhou *et al.* [26], but was limited by the fact that was applicable only to some types of CNNs. More general and effective methods are represented by GradCAM [21] and GradCAM++ [3]. They both are *gradient-based* methods that use the gradient (generated by the classification output of the network in a specific layer L) to produce the attention maps.

The concept of attention can be also partially exploited in unpaired image-to-image translation. In particular, Mejjati *et al.* [18] added an attention-guided generator to the CycleGAN architecture and used the attention as a way to separate foreground and background in order to apply the target domain to the former and not the to latter. Finally, Emami *et al.* [6] calculated attention in the CycleGAN discriminator and multiplied it with input image to guide the

generation. Both these approaches use the attention maps more as a mask than actively during the training as we propose. They are effective when a clear separation between foreground and background is present in the training samples, which is not always the case. For this reason, they are not designed to solve the cycle consistency limitation mentioned before.

3 Proposed Approach

Given two different domains X and Y, where $\{x_i\}_{i=1}^{N}$ are images belonging to X and $\{y_j\}_{j=1}^{M}$ are images belonging to Y, our model follows the CycleGAN formulation and therefore it is composed by two different generators G and F that learn the mapping $X \rightarrow Y$ and $Y \rightarrow X$, respectively. In addition to that, the model is also composed by two discriminators D_X and D_Y that learn to distinguish between real samples $\{x\}$ or $\{y\}$ and translated samples $\{G(x)\}$ or $\{F(y)\}$.

The baseline model is trained using an *adversarial loss* and a *cycle consistency loss*. The first one forces the generated samples to match the distribution of the target domain X or Y, while the second one allows to reverse the translation and avoids the generated samples to diverge from the input samples' shape. Nevertheless, cycle consistency loss introduces a drawback when the network tries to translate a domain X to a different domain Y or viceversa. Indeed, the cycle could prevent the generator to apply consistently the target domain in order to ease the reverse translation.

The main objective of this paper is to solve this drawback. This was achieved introducing a new loss term that forces the attention in the latent space of the two generators G and F to be the same when applying and reversing the translation, avoiding to lower the intensity of focus of the network during the cycle and denying the introduction of shortcuts when applying the target domain.

3.1 Attention Generation

Our method to calculate attention maps draws inspiration from GradCAM [21], but it introduces some modifications. In particular, since both D_X and D_Y follow the PatchGAN architecture [11], we compute the gradient by backpropagating the mean of the discriminator output $d = \frac{1}{P}\sum_p D(x)$ to the latent space of the corresponding generator and, more specifically, to the last layer of the last residual block $\mathbf{L}$. Global Average Pooling is then applied to the gradient to obtain the weight w_k:

$$w_k = \frac{1}{R}\sum_i \sum_j \left(\frac{\partial d}{\partial \mathbf{L}_k^{ij}}\right) \tag{1}$$

where $\mathbf{L}_k$ is the kth feature map with dimensions $w \times h$ of the layer $\mathbf{L}$ and $R = w \times h$. After this step, w_k is multiplied with the feature maps of the layer $\mathbf{L}_k$ obtaining the attention map $\mathbf{A}^d$:

$$\mathbf{A}^d = ReLU(\sum_k w_k \mathbf{L}_k) \tag{2}$$

where $ReLU$ is the rectified linear unit function.

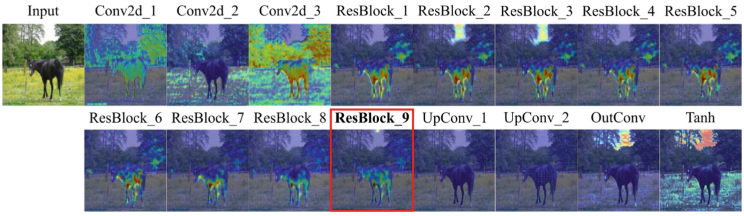


Fig. 2. Overview of attention maps extracted from the generator network at different layers. The proposed system uses the attention from the last residual block (in bold).

Finally, Fig. 2 shows all the attention maps generated from each layer of the generator network starting from an input image. Looking at the different attention maps, the one extracted from the last residual block is the one that precisely highlights the region where the domain needs to be applied. For this reason, the attention extracted from the last residual block is the one used in the loss that will be presented in the next section.

3.2 Network Architecture

Underlying Architecture. The proposed architecture is based on CycleGAN and therefore it is composed by two generators and two discriminators. More in detail, the generators are both composed by an encoder, a decoder and 9 residual blocks in between, while the discriminators follows the PatchGAN architecture introduced by Isola *et al.* in [11]. Indeed, an overview of the system is presented in Fig. 3.

During training, *adversarial loss* $\mathcal{L}_{adv}$ is used to push the generators G and F to produce realistic results belonging to the target domains X and Y,

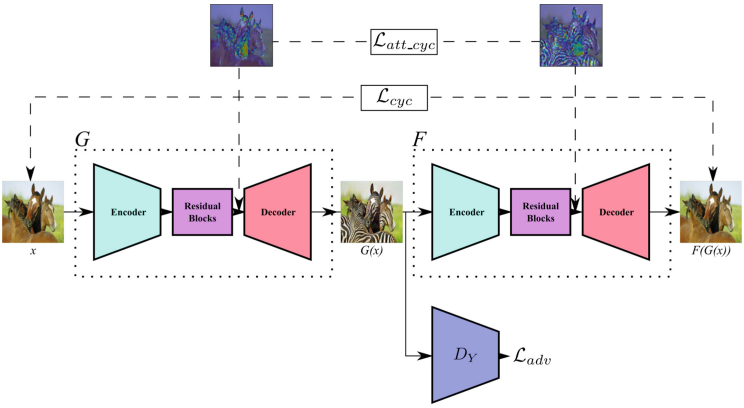


Fig. 3. Overview of the proposed system when translating an image from source domain X to target domain Y.

respectively, and *cycle consistency loss* $\mathcal{L}_{cyc}$ is used to force $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$ and $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$.

Attention Consistency Loss. The cycle consistency loss is very effective in avoiding the generation of undesired mapping during the translation, but, as stated before, it can also introduce shortcuts during training.

More in detail, in the generator's latent space the majority of the translation between source and target domains is performed with the encoder being mainly responsible of reducing the input dimension by encoding the image information, and the decoder that allows to go back to the original input shape. Nevertheless, during the decoding phase, the network can learn to reduce the intensity with which the domain is applied in order to facilitate the job of the inverse mapping generator. For example, an horse that was only half turned into a zebra would probably be considered a zebra by the corresponding discriminator, but it would be much easier to reverse it back to its original domain.

For this reason, we introduced a new term called *attention consistency loss* $\mathcal{L}_{att\_cyc}$ to improve the domain translation task. More specifically, the objective is to push the network to maintain the same attention over the whole translation cycle. Therefore, having $(x, F(y))$ and $(y, G(x))$ as input of G and F, respectively, the loss is:

$$\begin{aligned} \mathcal{L}_{att\_cyc} = \|\mathbf{A}^{D_Y}(x) - \mathbf{A}^{D_X}(G(x))\|_2 + \\ \|\mathbf{A}^{D_X}(y) - \mathbf{A}^{D_Y}(F(y))\|_2 \end{aligned} \tag{3}$$

where $\mathbf{A}^{D_Y}(x)$ and $\mathbf{A}^{D_Y}(F(y))$ are the attentions generated from the last residual block of G using the gradient obtained backpropagating from D_Y, while, similarly, $\mathbf{A}^{D_X}(y)$ and $\mathbf{A}^{D_X}(G(x))$ are the attentions generated from F backpropagating from D_X.

Imposing this new constraint during training helps the network to avoid any shortcut when applying the target domain, since the decoder will not dilute anymore the translation to ease the work of the cycle consistency loss term. The reason is that, in order to maintain the same level of attention in the two generators, the domain needs to be strongly applied to the source image without any compromise.

Finally, the full objective becomes:

$$\mathcal{L}_{D_X, D_Y} = \mathcal{L}_{adv} \tag{4}$$

$$\mathcal{L}_{G,F} = \mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{cyc} + \lambda_2 \mathcal{L}_{att\_cyc} \tag{5}$$

where λ_1 and λ_2 are set to 10 and 1, respectively.

Our system works without introducing any architectural change to the network. Therefore, the number of parameters of the proposed system are the same of Cyclegan, that is about 28.3 million.

4 Experiments

All experiments were executed training the network for 200 epochs using the Adam optimizer [13] with a learning rate of 0.0002. A qualitative and quantitative evaluation will also be performed. In particular, the latter has been done

using both the FID (Frechet Inception Distance) score [8] and the KID (Kernel Inception Distance) score [1].

4.1 Datasets

A subset of datasets used by CycleGAN have been selected for the experiments: *horse2zebra* (939 horse images and 1177 zebra images), *orange2apple* (996 apple images, and 1020 orange images), *photo2map* (1096 maps and 1096 aerial photos) and, finally, *monet2photo* (1074 Monet paintings and 6853 pictures).

Among all these, the last two datasets contain images where the translation process can not take advantage of a strict separation between foreground and background proving that our architecture will be effective also in these cases.

4.2 Results When a Foreground/background Separation Is Present

The objective of this set of experiments is to prove that introducing the attention consistency loss in the training has a positive effect over the results. In particular, it should solve the limitation of the cycle consistency loss that tends to maintain lots of features from the source domain in the translated image. In order to validate this claim, qualitative and quantitative results on the first two datasets will be presented.

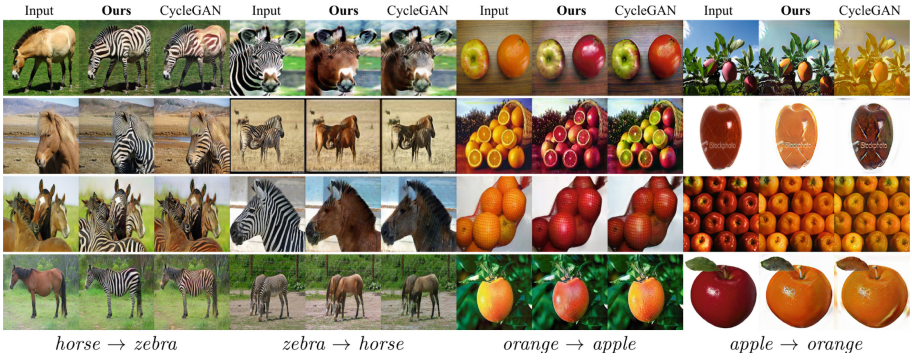


Fig. 4. Samples generated by our model wrt those generated by CycleGAN.

A comparison between samples generated by our model and sample generated by CycleGAN can be seen in Fig. 4. It is clear how results obtained with the aid of attention consistency loss are qualitatively superior to the ones obtained using a vanilla version of CycleGAN. In particular, CycleGAN is able to translate the domain somehow correctly, but the application is not consistent over the image. This is particularly evident in the *horse* → *zebra* domain transfer task where the stripes do not cover completely the original horse shape and most of the original color is still visible. On the other hand, when using the attention consistency

loss, the stripes are applied much more strongly and precisely over the animal body and almost no trace of the source domain is left. In addition to that, when doing the opposite transformation (which is much harder), the proposed system allows to remove stripes from the zebras more firmly and the overall translation is more convincing. On the other side, considering *orange* → *apple* and *apple* → *orange* translations, CycleGAN sometimes tends to left the original fruit almost unchanged or to apply a color filter over the whole image during the domain transfer, while after the application of the attention consistency loss the original fruit is not recognizable anymore in the image and no filter is applied. This leads to a sharper result.

Finally, a full quantitative evaluation has been carried out and the results are reported in Table 1. Our method outperforms CycleGAN in all the different domains, in terms of both FID and KID, proving that the introduction of the attention consistency loss is beneficial for the network.

Table 1. Quantitative results of our method compared with CycleGAN.

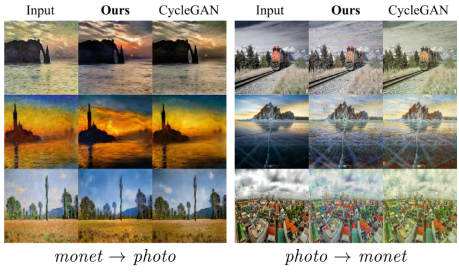
	FID ↓			
	Horse → *Zebra*	*Zebra* → *Horse*	*Apple* → *Orange*	*Orange* → *Apple*
CycleGAN [27]	33.66	64.57	105.15	81.05
Ours	**27.94**	**61.54**	**103.89**	**75.79**
	KID ↓			
CycleGAN [27]	0.013	0.026	0.058	0.042
Ours	**0.009**	**0.024**	**0.054**	**0.040**

4.3 Results When a Foreground/Background Separation Is Not Present

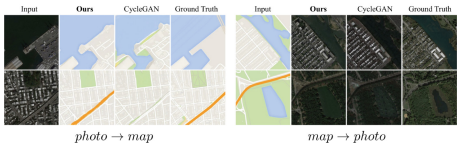
After proving the effectiveness of the attention consistency loss over domains where a clear separation between background and foreground was possible, we tested our system on datasets where there is no such separation.

Firstly, we experimented with style transfer and trained the network to produce images similar to a Monet painting starting from pictures, and vice versa. In this case, the translation needs to happen on the whole surface of the image and therefore methods like [18] would not be applicable.

Qualitative exemplar results of this task are shown in Fig. 5a. Indeed, it can be observed how, when translating from panting to photo, CycleGAN tends to maintain visible the brush strokes affecting the realism of the produced results, while our method produces more colorful and plausible results. In the opposite case, CycleGAN struggles with color variety, whereas our method is able to transfer the painter style much better. Finally, the improvement of our method wrt CycleGAN is confirmed by quantitative results (Table 1), where both FID and KID scores are lower (and therefore better) when applying attention consistency loss.



(a) *monet2photo* results.



(b) *photo2map* results.

Fig. 5. Some results with no foreground/background separation.

To further demonstrate the superiority of the proposed approach, we also experimented with map image translated in aerial view, and vice versa. Qualitative results for this experiment are shown in Fig. 5b. Indeed, our method is much better and precise than CycleGAN in reproducing water and highways on the maps images. Furthermore, when translating from maps to photos, it produces more realistic results than CycleGAN, especially for water and trees areas. These results were also validated by FID and KID values, reported in Table 2. CycleGAN has a slightly better FID score only in the case of the *photo* → *map* translation, but the corresponding KID score, a more reliable quality estimator, shows the effectiveness of our method.

4.4 Ablation Study

As mentioned before, the final setting of our architecture considers $\lambda_1 = 10$, $\lambda_2 = 1$ (see Eq. 5) and the generation of attention from the last residual block. This final setting has been obtained through an ablation study (performed using the *horse2zebra* dataset). Table 3 reports the results achieved with this study, where last line (row #6) corresponds to the final setting.

A first interesting experiment is to change the value of λ_2, while the value λ_1 has not been changed to be compliant with the original choice of CycleGAN. Increasing λ_2 to 10 (row #1) or decreasing it to 0.1 (row #2) do not bring to better results in terms of FID (results with KID are very similar in general). We also tried to apply the attention consistency loss $\mathcal{L}_{att\_cyc}$ only when translating the domain from X to Y and not when translating from Y to X (row #3) and only to F in the $X \rightarrow Y$ case and only to G in the $Y \rightarrow X$ case (row #4) in order to impose the loss only on the first generator in each cycle. Finally, the

loss was calculated extracting the attention from the last four residual blocks of the two generators instead of using only the last one (row #5). All these experiments lead us to the final setting mentioned above and reported in row #6, which achieves the best results.

4.5 Drawbacks

We have proved that when CycleGAN produces a result in which the translation is applied softly but correctly, our method will greatly boost the quality of the generated results. Nevertheless, there are some cases, like the ones presented in Fig. 6, where our attention transfer has the effect of enhancing the failure of CycleGAN in applying the target domain. For example, in the *horse* → *zebra* translation, if CycleGAN paints some stripes over the background our method could amplify it. Nevertheless, these effects rarely happen and only in some extreme cases.

Table 2. FID and KID results comparison between CycleGAN and the proposed method on *monet2photo* and *photo2map*.

	FID ↓			
	Monet → *Photo*	*Photo* → *Monet*	*Map* → *Photo*	*Photo* → *Map*
CycleGAN [27]	144.18	145.55	70.70	**63.61**
Ours	**141.90**	**140.69**	**55.55**	64.70
	KID ↓			
CycleGAN [27]	0.022	0.012	0.026	0.033
Ours	**0.019**	**0.011**	**0.013**	**0.025**

Fig. 6. Translation errors, already present in the CycleGAN output, that were enhanced by the attention consistency loss.

This limitation could be solved by combining the proposed method with ones like [18] where the translation is applied after separating foreground and background using the attention generated from an additional network. This is out of the scope of this paper and it is applicable only in the datasets like the ones in Sect. 4.2.

Table 3. Ablation study for different applications of the attention consistency loss.

		FID ↓	
		Horse → *Zebra*	*Zebra* → *Horse*
1	$\lambda_2 = 10$	30.66	62.69
2	$\lambda_2 = 0.1$	31.90	66.01
3	$\mathcal{L}_{att\_cyc}$ only $X \rightarrow Y$	30.60	65.32
4	$\mathcal{L}_{att\_cyc}$ single gen	32.74	62.71
5	$\mathcal{L}_{att\_cyc}$ 4 res blocks	33.18	61.77
6	**Ours**	**27.94**	**61.54**

5 Conclusions and Future Works

The objective of this paper was to cope with an important drawback of the cycle consistency loss used in unpaired image-to-image translation. In particular, this loss has the side effect of encouraging shortcuts when translating an image from a source to a target domain. The proposed solution exploits the attention maps extracted from the two generators of the network by introducing a new loss term called *attention consistency loss*. This loss forced the two generators to have the same attention in order to maintain the focus of the network high during the whole cycle.

Eventually, we proved the efficacy of the method by performing several experiments showing both qualitative and quantitative results, and in two main scenarios: the foreground and background clearly separated in the image, and scenarios where this separation is not present (typical cases of style transfer).

Future works will consist in testing the proposed loss to other architectures like [18] and also expand the use of attention maps to different tasks other than unpaired image-to-image translation.

Acknowledgments. This research has financially been supported by the Programme "FIL-Quota Incentivante" of University of Parma and co-sponsored by Fondazione Cariparma.

References

1. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd Gans. arXiv preprint arXiv:1801.01401 (2018)
2. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
3. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847. IEEE (2018)

4. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8789–8797 (2018)
5. Dhar, P., Singh, R.V., Peng, K.C., Wu, Z., Chellappa, R.: Learning without memorizing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019
6. Emami, H., Aliabadi, M.M., Dong, M., Chinnam, R.B.: Spa-GAN: spatial attention GAN for image-to-image translation. IEEE Trans. Multimedia **23**, 391–401 (2020)
7. Goodfellow, I.J., et al.: Generative adversarial networks. arXiv preprint arXiv:1406.2661 (2014)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. arXiv preprint arXiv:1706.08500 (2017)
9. Hoffman, J., et al.: Cycada: cycle-consistent adversarial domain adaptation. In: International Conference on Machine Learning, pp. 1989–1998. PMLR (2018)
10. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV), pp. 172–189 (2018)
11. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
12. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: International Conference on Machine Learning, pp. 1857–1865. PMLR (2017)
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 35–51 (2018)
15. Li, K., Wu, Z., Peng, K.C., Ernst, J., Fu, Y.: Tell me where to look: Guided attention inference network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
16. Liu, W., Li, R., Zheng, M., Karanam, S., Wu, Z., Bhanu, B., Radke, R.J., Camps, O.: Towards visually explaining variational autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8642–8651 (2020)
17. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5188–5196 (2015)
18. Mejjati, Y.A., Richardt, C., Tompkin, J., Cosker, D., Kim, K.I.: Unsupervised attention-guided image to image translation. arXiv preprint arXiv:1806.02311 (2018)
19. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
20. Nizan, O., Tal, A.: Breaking the cycle-colleagues are all you need. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7860–7869 (2020)

21. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
22. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)
23. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
24. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stack-GAN++: realistic image synthesis with stacked generative adversarial networks. IEEE Trans. Pattern Anal. Mach. Intell. **41**(8), 1947–1962 (2018)
25. Zhao, Y., Wu, R., Dong, H.: Unpaired image-to-image translation using adversarial consistency loss. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12354, pp. 800–815. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_46
26. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)
27. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)