I'm something of a Painter myself

Filipa M. M. Ramos
Faculty of Engineering of University of Porto

up201305378@fe.up.pt

Figure 1: Some paintings generated from photographs by the Encoder-Decoder architecture after tuning.

Abstract

The field of painting generation is a subset of image-to-image translation problems where the goal is to map an input image to an output image without presence of paired data. Focusing on the specific translation of photographs into Monet paintings, this work analyses the potentials of CycleGAN on these specific domains. Some improvements to the base model are proposed, including an encoder with shared weights, a tuned architecture and an asymmetry enforcing parameter. These configurations are explored using both qualitatively and quantitatively evaluation methodologies, showcasing significantly better results than the initial baseline.

1. Introduction

Image-to-image translation is the field of computer vision that studies the learning of a mapping between an input and output image. Many applications can be identified for this problem, starting from, but not limited to, collection style transfer, object transfiguration, season transfer and photo enhancement.

Learning a mapping between an input and output image is a process that encompasses several challenges of interest. Firstly, how can we define the output in a rigorous and scientific way? The lack of definition of the output is common. Rather, the output is many times subject to interpretation, such as the case of painting generation [19]. Secondly, how can this task be successfully formulated in an artificial system? How to direct the learned mapping towards the expected output? And, just as important, how to objectively evaluate the quality of the translation? Finding a solution to these questions could have a heavy impact on a wide range of scientific areas including image processing, medicine and linguistics just to name a few.

Even though the area of image-to-image translation in itself is of high interest to the scientific community, there are several demanding types of challenges faced in these studies. The existence of paired data, for a direct definition of the target, is many times impossible to obtain due to high costs of collection or the inability to even find such samples in nature. For this reason, it is of interest to study these classes of problems using unpaired data, modifying the premise in order to learn a mapping between an input and output image that is not comparable to a direct pair but rather belonging to the target set.

This work focuses on a specific field of image-to-image translation. Leveraging unpaired data, the objective is to obtain a mapping of photographs into what could be Monet paintings. The data and a suitable evaluation method are sourced from a Kaggle challenge entitled *I'm Something of a Painter Myself* [8].

The field of painting generation presents difficulties of its own. Some of the initially considered questions, upon analysis of the data, are related to the distribution of the target set. In fact, Monet presents paintings with a high variance in style - some look almost blurred, with less definition in the brush strokes whilst others look almost like photographs. Some feature dark, strong colours whilst oth-

ers are fuzzy all over. A perfect translation could perhaps only be achieved when and if we can transfer artistic and subjective concepts to machines. The reversed translation showcases a much more well defined target as photographs all have the same visual characteristics, changing more of the content rather than the visual style.

Specifically, the starting point of this work will be CycleGAN [19], due to the fact that it has been one of the most prominent solutions since its first publication in 2017. Even though the model is not particularly recent, there have been few to none improvements proposed. For this reason, the starting point of this work will be the original model proposed by Zhu *et al.* The expected learning outcomes can be highlighted:

- Creation of a working CycleGAN model [19] using Pytorch.
- Analysis of possible improvements for the CycleGAN model on the field of painting generation.
- Experimentation with different ideas to gain insight into the workings of the model.

2. Related Work

Paired image-to-image translation Pix2pix [7] is one of the major references for supervised image-to-image translation. Conditional GAN's [13] are used to learn both the mapping and loss function automatically. The model is applicable to a wide range of domains without overhead since the loss function does not need to be tuned.

Unpaired image-to-image translation The baseline model of this work, CycleGAN [19], proposes the addition of cycle consistency in order to restrain the set of possible output transformations. The concept is based on the idea that if X is translated to Y, Y can be translated back to X. To fulfil this, the authors leverage two generators, $G : X \longrightarrow Y$ and $F : Y \longrightarrow X$ and the corresponding discriminators D_X and D_Y. The accuracy of reconstruction is enforced through the addition of a cycle consistency loss.

Other works assume a shared latent space [11, 6]. The hypothesis of the shared latent space is based on the idea that images from different domains can be mapped to the same space by a variational auto-encoder [9]. These systems usually comprise two different GAN's [12], one for each domain. MUNIT [6] expands this idea by having a content and a style space.

More recently, contrastive learning [14] has proposed maximising the information between patches of the input and output through the sampling of unrelated patches that are used as negative examples. These are drawn from the input image itself, minimising overhead.

Style Transfer One of the original approaches, proposed in [2], makes use of the intermediate layers from an object detection network, VGG19 [17], in order to transform an image into a reference style and content. The features extracted by VGG19 are used to approximate the output to both the content, using the mean squared error, and the style, using gram matrices.

3. Method

The goal in sight is to learn a mapping from a domain X to a domain Y. Depicting the generator as G, our objective is then to learn $G : X \longrightarrow Y$. The discriminator that is tasked with distinguishing the generated Y from a real sample is represented as D_Y. The adversarial loss, $\mathcal{L}_{GAN}$, in 1, enables the equilibrium learning for both networks [3].

$$\begin{aligned} G* = \arg min_G max_{D_Y}\ & \mathbb{E}_{y\sim\mathbb{P}_{data}(y)}[\log D_Y(y)] \\ & + \mathbb{E}_{x\sim\mathbb{P}_{data}(x)}[\log(1 - D_Y(G(x)))] \end{aligned} \tag{1}$$

The negative log-likelihood on this loss might cause gradients to become infinite which leads to unstable training procedures. For this reason, the mean squared error is also considered for adversarial training, with modified parcels $\mathbb{E}_{y\sim\mathbb{P}_{data}(y)}[(D(y) - 1)^2]$ and $\mathbb{E}_{x\sim\mathbb{P}_{data}(x)}[[(D(G(X)) - 1)^2]$.

CycleGAN proposes the addition of two terms to the target loss function in order to restrain the output space and guarantee that the learning is successful. Considering generators $G : X \longrightarrow Y$ and $F : Y \longrightarrow X$, the cycle consistency, in 2, reinforces the reconstruction's similarity to the original image, i.e., $F(G(X)) \sim X$ and $G(F(Y)) \sim Y$.

$$\begin{aligned} \mathcal{L}_{cyc}(G, F) = & \mathbb{E}_{x\sim\mathbb{P}_{data}(x)}[\|F(G(x)) - x\|_1] \\ & + \mathbb{E}_{y\sim\mathbb{P}_{data}(y)}[\|G(F(y)) - y\|_1] \end{aligned} \tag{2}$$

Further than this, an identity term is also considered in order to discourage the network from changing the original tints of the image too much 3.

$$\begin{aligned} \mathcal{L}_{id}(G, F) = & \mathbb{E}_{x\sim\mathbb{P}_{data}(x)}[\|F(x) - x\|_1] \\ & + \mathbb{E}_{y\sim\mathbb{P}_{data}(y)}[\|G(y) - y\|_1] \end{aligned} \tag{3}$$

The full CycleGAN objective can then be expressed as 4. The cycle consistency and the identity term are weighted in order to further tune their impact on the loss function.

$$\begin{aligned} \mathcal{L}_{CycleGAN}(G, F, D_Y, D_X) = & \mathcal{L}_{GAN}(G, D_Y, X, Y) \\ & + \mathcal{L}_{GAN}(F, D_X, Y, X) \\ & + \lambda\mathcal{L}_{cyc}(G, F) \\ & + \lambda_{id}\mathcal{L}_{id}(G, F) \end{aligned} \tag{4}$$

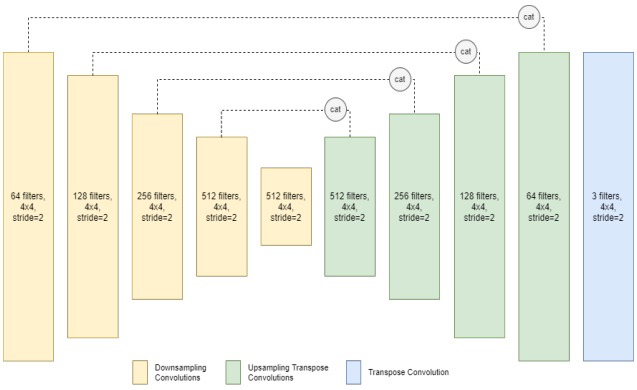

Figure 2: Best performing architecture for the generator. Based on the U-Net [15] model, the output of the downsampling layers is concatenated with the corresponding upsampling layers.

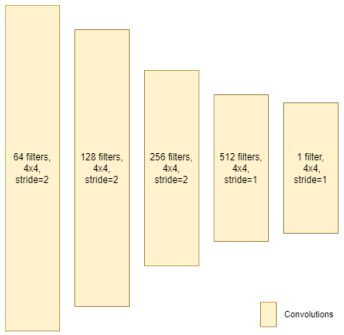

Figure 3: Best performing architecture for the discriminator is very close to the original Patch discriminator with slight adjustments on the final layers.

4. Implementation

The full implementation is achieved with the help of the PyTorch framework for Python. The initial model, as a baseline, was constructed based on the CycleGAN paper, leveraging a U-Net based generator and a pixel discriminator. Further than this, two other versions were implemented. The first was a complex encoder-decoder style generator paired with a deeper discriminator. Noise sampled from a normal distribution with mean 0 and variance 1 is added to the output of the encoder before feeding it to the decoder. Residual blocks are present in both the encoder and decoder. Transposed convolutions are not performed on this version, and nn.UpSample is used instead. The other version was based on the original U-Net generator with a more tuned architecture and a discriminator with 5 convolutions. Contrary to the previous version, transposed convolutions are used for the upsampling.

Following the original paper, discriminators are fed a random sample from a buffer of 50 previously generated images, the learning rate is linearly decayed in the last epochs and the batch size is kept as 1.

4.1. Network Architectures

For simplification, only the best performing architectures are illustrated. The final generator architecture is presented in figure 2. The model features 5 downsampling convolutions with instance normalisation, reflection padding and a leaky relu activation. Through experimentation, it was observable that removing the normalisation from the first layer increased results significantly. The upsampling is achieved with transpose convolutions including instance normalisation and a relu activation. All convolutions performed maintain a filter size of 4x4 and a stride of 2.

The discriminator's architecture is very close to the original Patch discriminator, with slight adjustments. Figure 3 showcases the full architecture. The model is composed of 5 convolutions followed by instance normalisation (except on the first layer) and leaky relu activation.

5. Experiments

The dataset available for experiments comprises 300 Monet original paintings and 7038 random photographs. The photographs are mainly representative of natural landscapes with a few featuring people and strongly contrasted scenes. Upon careful inspection of the set of Monet paintings, it becomes apparent that there is a wide range of styles and techniques in between samples. In fact, the distribution seems hard to describe, even in natural language.

Numerical evaluation is performed through Kaggle's platform. Their evaluation methodology uses the Memorization-informed Fréchet Inception Distance, short MiFID, which is a variation of FID [5] that takes into account training sample memorisation. The FID [5] between real and generated images is calculated using the extracted features from an intermediate layer of an Inception network [16]. The features are modelled as a multivariate Gaussian distribution and the Fréchet distance expresses the distance between the distributions of features of the real and generated images. The FID is then divided by the minimum cosine distance between all training samples in feature space, which composes the final MiFID score. The lower the MiFID score, the better is the general quality of the translations. Further evaluation is performed through visual analysis of the intricacies of generated images.

5.1. Experimental Variations

As previously mentioned, several different architectures were tested for both the generator and discriminator. A history of experiments is summed up for each specific model.

Figure 4: Some of the samples from the first split of Monet paintings. These paintings showcase consistent colours and a similar brush technique.

Model	MiFID score
Baseline	71.89
Encoder-Decoder	62.95
Improved baseline	**44.77**

Table 1: Best MiFID scores obtained for each implemented version.

Baseline architecture The initial parameters were based on the usually recommended GAN procedures, including a leaky relu activation with a slope of 0.2, batch normalisation, a 2x2 stride for downsamples and the adam optimisation methodology with a learning rate of 0.0002 and momentum of 0.5. The adversarial loss considered for the baseline is the mean squared error. After some optimisation, a MiFID score of 71.89216 is obtained.

Encoder-Decoder architecture During the training process of these experiments, a steep increase in GPU memory consumption was noted, with a batch size of 1 using around 4.5GB of memory. Moreover, training time exponentially increased, from around 4 to 12h. With some optimisation, and using the negative log likelihood loss in 1, the evaluation outcome is a score of 62.94714. Figure 1 showcases paintings generated with this architecture.

Improved baseline architecture Several experiments were carried out using different λ and λ_{id} values. λ_{id} was varied in the interval $[0, 5]$. With a λ_{id} of 2, the best score was achieved, 44.76828. The extension of the generator from 5 to 7 layers was also tested, however, results deteriorated slightly.

Further than the tuning of the several implemented variants, table 1, some additions to both the encoder-decoder architecture and the improved baseline were considered:

- Injection of noise sampled from a normal distribution in the discriminators. The premise is that this could break the initial training advantage of the discriminators in face of the generators.
- Varied data augmentation settings. Data is augmented in place and randomly through the training epochs.
- Improved WGAN [1, 4] loss strategy and gradient penalty for training stability and result refinement.
- Creation of sub-datasets of Monet paintings that present more similarity in style between each other.
- Shared weights between the Monet and photo encoders. The idea is to indirectly enforce the shared latent space assumption [11, 6].
- Addition of a penalisation term to the Monet discriminator loss. Destined to create asymmetry in training objectives. Loosely based on [10]. Under the premise that $Monet \longrightarrow photograph$ could be considered a more defined translation than $photograph \longrightarrow Monet$.

(a) Translation with red spots. Obtained with parameters $\lambda_{id} = 0.2$ and $\lambda_{cyc} = 10$.

(b) Translation with black spots. Obtained with parameters $\lambda_{id} = 0.4$ and $\lambda_{cyc} = 10$.

(c) Translation with black spots. Obtained with parameters $\lambda_{id} = 0.5$ and $\lambda_{cyc} = 10$.

Figure 5: Artefacts commonly seen with data augmentation during training. The variation of the weight given to the identity loss makes the artefacts more prominent.

Figure 6: AMT perceptual tests given to the participants. Each pair has a real Monet painting and a generated one.

As for the injection of noise, no clear improvements were seen. For the data augmentation study, however, some interesting results were obtained. Alternating random crops, flips, perspective wraps and gaussian blur lead to the appearance of strange artefacts in the translated images 5. The only data augmentation techniques that did not lead to the appearance of these were the flips, both horizontally and vertically. Even though these did not lead to deformations in the output, impact on performance was not noticeable.

The addition of gradient penalty and the adequate adaptation of the adversarial loss, following [4], demonstrated a lot more training stability, however, the model could not converge to the desired output, degenerating to reconstructing the input images. This behaviour was equally observed with the identity term and without it.

The creation of sub-datasets was carried out through splitting the original set of paintings into two sub-sets, with 153 and 147 respective samples. The first set contained only paintings with a more defined brush stroke pattern, consistent colours and subjects - these could be considered in the style that Monet is most known for, figure 4. The other set showcased much more varied painting styles, with either strong or faded colours and more exotic scenes. Experiments on these sub-sets of paintings did not seem to have a significant impact on the translation outcomes, neither positive nor negative.

Furthermore, using the Encoder-Decoder and the improved baseline architectures, a shared weights approach was tested. The shared weights were respectively applied to the encoder and/or decoder and the downsampling and/or upsampling layers. The best performing experiments were with shared weights only on the encoding layers, i.e., the encoder and the downsampling. Comparing the same configuration with and without the shared weights setup, even though the MiFID score worsened slightly, visually, it was observable that the shared weights setup produced consistently paintings with less overall noise and better scene definition, figure 7.

Finally, the asymmetry hypothesis was tested directly in the Monet discriminator's loss with a penalisation of 0.5 and 0.3. The averaged values of the real and generated discriminator loss values were multiplied by this term. Using a penalisation of 0.5, the equilibrium is too strongly broken which lead to worse translations and less training stability. However, with a penalisation term of 0.3, the results showcased great visual accuracy, even though the best obtained score was worsened slightly, and the training stability of the Monet generator and discriminator was benefited.

AMT Perceptual Study A small perceptual study [18] was also conducted in order to ascertain the visual quality of the generated paintings. With a sample of 4 participants, this study provides only a reference of the common visual accuracy of the obtained translations. The experimental setup consisted of 10 tests with pairs of one original Monet painting and a generated one. The participants were asked to choose, for each pair, which they thought was the real painting and the generated one. Figure 6 represents the tests that were given to the participants.

On average, participants achieved 77,5% of accuracy in choosing the correct Monet painting. The last test was the one that proved most difficult for participants with only 50%

Figure 7: Some visual differences found in the translations made by the without/with shared weights configurations, respectively from left to right.

Identity Loss Configuration	MiFID score
present, $\lambda_{id} = 2$	**44.77**
not present	51.71

Table 2: MiFID scores for the identity loss ablation.

being able to point out the real Monet painting. Participants with previous knowledge of Monet's works were more efficiently able to point out the real paintings, however, they still described the test as difficult and challenging.

5.2. Ablation Study

The original CycleGAN paper [19] presents an ablation study in regards to the forward and backward consistency parcels of the loss function. However, since the identity loss is only used for painting generation, the authors did not include this parcel in the analysis. For this reason, it is of interest to analyse how impactful the presence of the identity loss term is to the model's performance.

In order to fulfil this, two experiments were carried out, using the previously optimised parameters. For these experiments, it is important to take into consideration two different metrics, MiFID score and training stability. Table 2 denotes the contribution of the identity loss to the model's accuracy. Further than this, it was also observed that the removal of the identity parcel made the training process more unstable, figure 10, with the adversarial loss in particular unstably rising through time.

Through a visual analysis of the obtained paintings, many good results are found, as showcased on figure 8. Even though the translations seem positive, there is more noise found in the output when the term is removed. Moreover, figure 9 reveals strong negative examples on which it is possible to denote that the model makes some drastic colour changes when the identity term is absent.

6. Discussion

Concerning the data augmentation artefacts, a reasonable explanation for this could be the fact that the transformations are applied randomly through training. This means that the target distribution will be changed during learning by a stochastic process which might cause the degeneration into the seen anomalies. Furthermore, since the distribution is changed suddenly, the next output from the generator can be classified as being fake even if its translation is of high quality which leads to higher loss values for the generator that can deteriorate its optimisation path. Horizontal and vertical flips change the content and the style the minimum and should, in theory, make the modelling of the distribution rotation invariant. It is not clear whether this could be beneficial or not. The style that is trying to be replicated describes natural scenes with minimal changes in colour. It is arguable whether a perfectly translated, however vertically flipped image should be classified as a Monet painting or not.

The results obtained with the sub-datasets were below expectations. Several configurations trained on the defined set of paintings showcased similar results to the training on the entire set. The brush strokes found on the originals were still not present on the translation outputs which might indicate that the discriminator is not using these visual details in classification. An interesting improvement of the discriminator would be to enforce its attention on smaller patches that detailed the painter's techniques. Overall, the obtained paintings seem to indicate that the discriminator is effective in classifying images as being paintings or not, however, it is not capable of enforcing concretely the style of Monet. Figure 1 describes this well. We can certainly classify these translations as paintings. Could we classify them as a work of Monet?

Forcing the shared weights configuration might not lead the encoding of both domains into the intended shared latent space. The encoding of the different domains might even have isolated semantic meanings, independently of the encoder weights. This could explain the lack of an obvious improvement in performance. Further than that, the learning optimisation could lead to similar weights on both encoders even without these being forced since the reconstruction loss has a great impact on the learning process. Pairing a generator-discriminator after the shared weights encoders could force the mapping to a shared space and possibly improve results.

As for the asymmetry hypothesis, more testing is needed in order to ascertain its true impact. However, asymmetry has been proven to be beneficial for GAN training and the idea that a $photograph \longrightarrow Monet$ translation carries a lot more information than the opposite translation might indicate that the learning objective is asymmetric in itself. CycleGAN indirectly carries the idea that a translation $X \longrightarrow Y$ and $Y \longrightarrow X$ have the same cost. One could argue that turning a painting into a photograph is a much more defined translation and even easier to perform than the opposite one.

Figure 8: Some positive examples observed without the identity loss term. Even though the translations are fairly good, the instability of the training is showcased through the added noise.

Figure 9: Negative examples are more prominently observed without the identity loss term. The lack of the loss term for the consideration of the photograph's identity leads to strong colour changes and added noise to the translation.

Figure 10: Removing the identity loss makes training unstable and breaks the generator-discriminator balance. Generator losses start increasing after a few epochs.

7. Conclusion

The main objective of this work was to identify possible improvements for the CycleGAN model in the context of a direct translation from a photograph to a Monet painting. Further than this, it was of high importance to study the visual quality of the painting generation and the training procedure of the GAN's.

In order to fulfil this, different architectures were tuned towards better results. Using the optimised networks, several hypothesis were tested as additions to the improved CycleGAN model. Furthermore, an identity loss ablation study was provided in order to give insight into the inner workings of the model, especially its training procedure.

Having improved massively upon the initial baseline score, it is considered that some relevant and efficient improvements were proposed. A thorough analysis of the target paintings and the generated ones demonstrate that the employed methodology is able to produce good results. In fact, it might even be possible to fool human testers in some cases.

However, the modelling of the target distributions lacks the attention to intricacies of artistic quality. Further than this, it was noted during the perceptual tests that the main reason for identification of the generated painting was the slight noise that is consistent through all the translations, especially in the corners of the images. The inclusion of an image quality term to the loss function might be an interesting open point of investigation. Moreover, the pursuit of the idea of asymmetry in the painting generation setting can be a study of great value for the field. Other interesting ideas would be to introduce negative examples from paintings of other painters.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 2017.

[2] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style, 2015. cite arxiv:1508.06576.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc., 2014.

[4] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5767–5777. Curran Associates, Inc., 2017.

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 6626–6637. Curran Associates, Inc., 2017.

[6] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018.

[7] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

[8] Kaggle. I'm something of a painter myself, 2020.

[9] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1558–1566, 2016.

[10] Yu Li, Sheng Tang, Rui Zhang, Yongdong Zhang, Jintao Li, and Shuicheng Yan. Asymmetric gan for unpaired image-to-image translation. *IEEE Transactions on Image Processing*, PP:1–1, 06 2019.

[11] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 700–708. Curran Associates, Inc., 2017.

[12] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 469–477. Curran Associates, Inc., 2016.

[13] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2642–2651, 2017.

[14] Zhang R. Zhu JY. Park T., Efros A.A. Contrastive learning for unpaired image-to-image translation. *Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020.*, 12354, 2020.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, May 2015.

[16] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 2234–2242. Curran Associates, Inc., 2016.

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[18] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016.

[19] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.