Taking off and putting on eyeglasses using Cycle-Consistent Adversarial Networks

Project report

Mariia Stazherova

Supervisors: *Dr. Paul Prasse, Prof. Dr. Tobias Scheffer*

University of Potsdam
April 8, 2019

Contents

Abstract

The purpose of this research project is to apply cycle-consistent adversarial networks proposed by Zhu et al. [1] to the face attribute manipulation problem. In the project we investigate the use of unpaired image-to-image translation to the task of automatic eyeglasses removal from frontal facial images along with the reverse task of adding eyeglasses to facial images. The results on the training set show high potential for both tasks, but not so successful generated images on previously unseen test set indicate that the framework still needs fine-tuning and improvement.

---

1 Introduction

Over the last decades, face recognition has become a major research area in computer vision and image analysis due to its interdisciplinary nature and numerous application domains, from computer science to neuroscience and psychology. Examples of practical applications of face recognition systems include biometric identification, information security, surveillance, multimedia and entertainment, diagnosing diseases, access control and many others. Despite an impressive amount of successful methods, face recognition still remains a challenging task since its performance is significantly affected by different face variation factors, such as lighting, pose, facial expression and occlusions.

Eyeglasses are one of the most common sources of natural occlusions and many deep learning approaches have been proposed to handle glasses-occluded face images. Liang et al. (2017) trained deep convolutional neural networks to learn the mappings between pairs of face images with and without eyeglasses. The approach effectively removes the eyeglasses region and reconstructs this region by selecting one piece from the facial image [2]. Generative adversarial networks (GAN) [3] are usually used for the face attribute manipulation problem (which aims at modifying a face image according to a given attribute value) because it is very difficult to collect labeled data for this task. Shen and Liu (2017) propose to learn a residual image defined as the difference between images before and after the manipulation instead of manipulating the whole image. This work shows effective transformations for several attribute manipulations, including removing or adding eyeglasses, closing or opening the mouth [4].

Motivated by the general-purpose property of cycle-consistent adversarial networks (CycleGAN), introduced by Zhu et al. in 2017 [1], and their ability to translate between domains without paired input-output examples, in this project we adapt a

CycleGAN to the task of removing eyeglasses from faces and the reverse task of adding eyeglasses to faces.

Eyeglasses are commonly divided in three categories: thin eyeglasses, thick eyeglasses, and sunglasses. Following Guo et al. [5], for this project we only focus on thick black-framed eyeglasses, since the impact of thin eyeglasses is too small whereas that of sunglasses is too big because of the severe identity information loss.

2 Related work

Image-to-image translation is the task of learning to map images from one class to another, e.g. colorization, style transfer, super-resolution, edge-map to photograph translation, semantic segmentation. It aims at transferring images from the source domain to the target domain by preserving domain-independent features while replacing domain-specific features.

Generative adversarial networks were introduced by Goodfellow et al. in 2014 [3] and their applications are still rapidly increasing. GAN is an unsupervised framework and the main idea behind it is having two neural networks, a generator and a discriminator, compete against each other. The discriminator tries to distinguish real data from fake data produced by the generator, while the generator's objective is to turn random noise into images that will be good enough to fool the discriminator. The generator uses the discriminator's score as its feedback mechanism to improve and to learn the underlying distribution of real data. Due to this opposing competition, the generator and discriminator are known as adversaries, and this method of training is known as the min-max game, where the discriminator minimizes a cross-entropy, but the generator maximizes the same cross-entropy. GANs have proven to be one of the most powerful tools for manipulating and generating images.

Pix2pix method designed by Isola et al. (2017) for paired image-to-image translation uses a conditional GAN (cGAN) to learn a mapping from a source image to a target image. Conditional GANs not only learn the mapping from input image to output image, but also learn a loss function to train this mapping. In a cGAN the generator gets an image as input instead of random variables and as long as the input is sufficiently complex it can play the role of noise. Instead of an encoder-decoder architecture for the generator, the authors used a U-Net [6], which includes skip connections directly connecting encoder layers to decoder layers. Isola et al. used a convolutional PatchGAN discriminator, which only operates over local regions of the image [7].

CycleGAN Building on the pix2pix framework, Zhu et al. (2017) address the problem of finding paired data and the issue of mode collapse by adding a cycle consistency loss to constrain image translation output to contain much of the information of the input. [1]. We apply this framework of cycle-consistent adversarial networks to our face attribute transformation problem (removing/adding eyeglasses) and describe detailed model's workflow in the following section.

3 Method

3.1 Task definition

In this section we will briefly describe the objective of a CycleGAN as defined by Zhu et al. [1] in view of the fact that we did not change it for this project. The overall goal is to obtain mappings between two domains, given unpaired image samples $\{a_i\}_{i=1}^{N}$ and $\{b_j\}_{j=1}^{M}$, where $a_i \in A$, $b_j \in B$, N is the number of samples for the domain A and M for the domain B. The model includes mapping functions in both directions: $G_b : A \rightarrow B$ (the generator G_b learns a mapping from the domain A to the domain B, e.g. faces with no glasses to faces with glasses), and $G_a : B \rightarrow A$ (the second generator G_a learns the reverse mapping function, e.g. faces with glasses to faces with no glasses). Two adversarial discriminators D_a and D_b attempt to predict if samples come from the actual distribution ($a \in A$ or $b \in B$ respectively) or produced by the generator ($G_a(b)$ or $G_b(a)$ respectively).

Adversarial loss ensures that the generator produces samples which are identical to the target distribution. Following the usual GAN [3] loss, the objective for $G_b : A \rightarrow B$ is expressed as:

$$\mathcal{L}_{GAN}(G_b, D_b, A, B) = \mathbb{E}_b[logD_b(b)]+ \\ \mathbb{E}_a[log(1 - D_b(G_b(a))],$$

where $b \sim p_{data}(b)$ and $a \sim p_{data}(a)$. Generator G_b aims to minimize this objective, whereas the discriminator D_b aims to maximize it, creating the so-called "min-max game". $\mathcal{L}_{GAN}(G_a, D_a, A, B)$ for the second mapping function is defined in the same way.

Cycle consistency loss was introduced in CycleGAN to be able to train the model given unpaired data. This loss can be viewed as a form of regularization because it reduces the space of possible mapping functions. The idea behind cycle consistency is that if we translate an image from the domain A to the domain B and then back to the domain A again, we should get the original image, i.e. $a \rightarrow G_b(a) \rightarrow G_a(G_b(a)) \approx a$. Zhu et al. call this translation forward cycle consistency, and similarly, the cycle $b \rightarrow G_a(b) \rightarrow G_b(G_a(b)) \approx b$ backward cycle consistency. Cycle consistency loss for the model

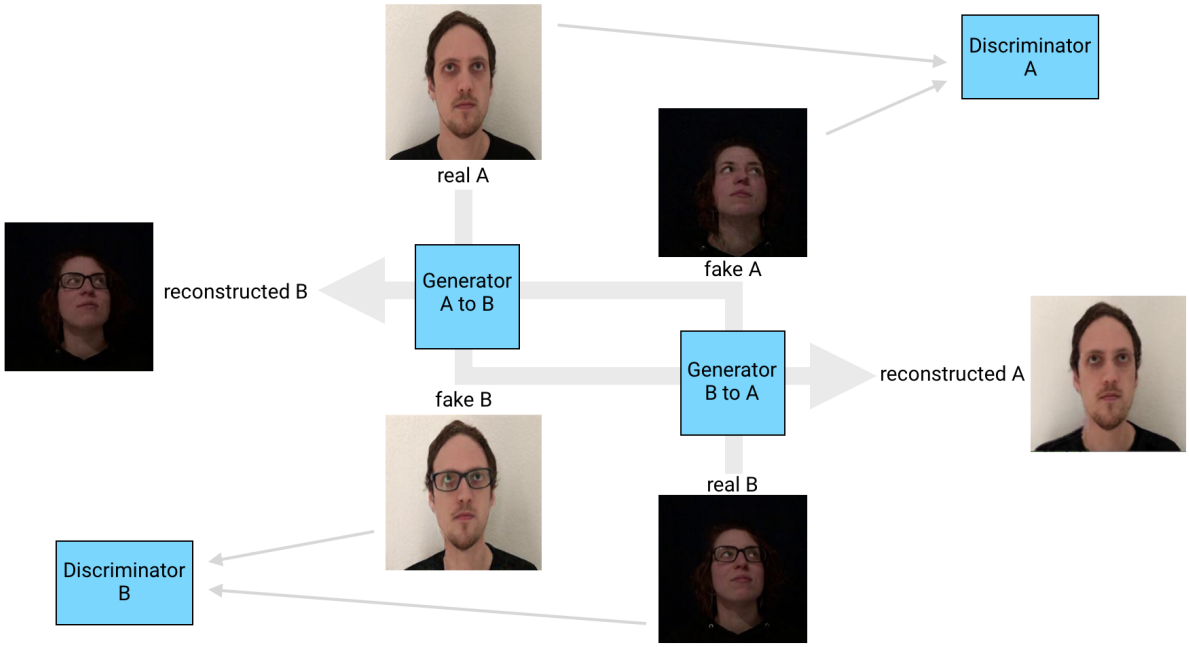


Fig. 1: The complete architecture of the model.

is thus formally expressed as:

$$\mathcal{L}_{cyc}(G_b, G_a, A, B) = \\ \mathbb{E}_a[\|G_a(G_b(a)) - a\|_1] + \mathbb{E}_b[\|G_b(G_a(b)) - b\|_1],$$

where $a \sim p_{data}(a)$ and $b \sim p_{data}(b)$. Cycle consistency prevents generators from excessive hallucinations and mode collapse and encourages them to generate images that share structural similarity with inputs.

Full objective of a CycleGAN is formed combining adversarial and cycle consistency losses together:

$$\mathcal{L}_{cyc}(G_b, G_a, A, B) = \mathcal{L}_{GAN}(G_b, D_b, A, B) + \\ \mathcal{L}_{GAN}(G_a, D_a, A, B) + \lambda \mathcal{L}_{cyc}(G_b, G_a, A, B),$$

where λ parameter assigns more importance to the cycle loss.

3.2 Model architecture

We adapted the original network architecture of the CycleGAN proposed by Zhu et al. [1] with some minor changes. Simplified view of the network is shown in Figure 1. The model contains two generator networks and two discriminator networks. Each generator-discriminator pair learns a mapping from source to target domain, following the principle of the original Generative Adversarial Network (GAN) first introduced by Goodfellow et al. [3].

For both discriminator networks, we used the same architecture as Zhu et al. [1], specifically a 70×70 PatchGAN, a convolutional classifier that penalizes structure at the scale of overlapping image patches as described by Isola et al. [7]. The output of the discriminator is a 30×30 image where each pixel value represents how believable the corresponding section of the unknown image is. Each pixel from this output image corresponds to the likelihood of a 70×70 patch of the input image. The detailed structure of the discriminator can be seen in Figure 10 in Appendix.

The architecture of both generator networks are the same and reproduce a U-Net, a popular end-to-end encoder-decoder fully convolutional network, originally designed for biomedical image segmentation [6]. Although ResNet models with several residual blocks were used in the original CycleGAN, the U-Net showed a slightly better performance for our dataset. The structure of a U-Net consists of two paths: an encoder, which downscales the image to a latent space using convolutional layers (also called contractive path); and a symmetric decoder, which upscales the bottleneck layer back to the original dimensions using transposed convolutions (also called expansive path). Skip connections between the encoder and the decoder apply a concatenation operator and provide better feature information. Figure 11 in Appendix shows the detailed generator architecture.

3.3 Training details

We train the model for 150 epochs which takes around 30 hours on the Nvidia GeForce GTX TITAN X graphics card, kindly provided to us by the Institute of Computer Science at the University of Potsdam. Following the original hyperparameters

proposed by Zhu et al. [1] we use Adam optimizer throughout the training with $\beta 1 = 0.5$, $\beta 2 = 0.999$ and $\eta = 2 \times 10^{-4}$. The initial learning rate is kept fixed for 100 epochs and is linearly decayed to zero for the remaining epochs. The batch size is set to one. Since four networks (two generators and two discriminators) have to be loaded on one GPU, training CycleGANs is a memory intensive task and it is one of the reasons Zhu et al. [1] chose the batch size of one, which allowed to train the model on images with higher resolution. Another reason is the use of instance normalization over batch normalization, which has an effect of making the output invariant to mean and variance of each feature channel of the input. Previous work [8] [9] has proven instance normalization to be successful for the task of style transfer and it has also been used as a replacement for batch normalization in GANs.

To stabilize the training, we update the discriminators with a history of 50 most recent generated images, instead of one. This technique was pioneered by Shrivastava et al. [10] and is widely used for reducing model oscillations in GANs [11] [1]. We use 50 as a size of the image buffer to follow the original settings by Zhu et al. [1]. To slow down the rate at which the discriminator learns, we also divide its objective by the factor of two. We update the generative and discriminative models asynchronously. Cycle loss lambda is set to 10. When training the discriminator, instead of providing 1 and 0 labels, we use soften values such as 0.9 and 0.1, which is shown to make the classifier more resilient to label noise.

Training set is shuffled every epoch. The weights of convolutional filters are initialized from a Gaussian distribution with mean 0 and standard deviation 0.02. All networks were implemented with Keras and the complete source code is available on Github under https://github.com/mstazherova/IM-CycleGAN.

4 Experiments

4.1 Data

Dataset The dataset contains multi-view head shots extracted from twelve RGB video sequences of two individuals (one male, one female). The video sequences were recorded with an iPhone X at a resolution of 1920×1080 pixels and a frame-rate of 60 frames per second. Each sequence consisted of a person rotating their head continuously dropping the chin toward the chest looking down, circling the head to one side, continuing the rotation looking up and to the opposite side, then down and back to the center. For each individual, videos were taken varying facial details (glasses / no glasses) and backgrounds (4 different settings) and their length ranged from 17 to 24 seconds. Frames were extracted from videos every 60 milliseconds, cropped to the same size of 256×256 pixels and saved as a sequence of JPEG images. More detailed distribution of the resulting images for each individual and setting is shown in Table 1.

Fig. 2: Training examples: eyeglasses (first row) and no-eyeglasses (second row) sets.

	Set 1	Set 2	Set 3	Set 4
X glasses	358	405	349	302*
X no glasses	368	395	280	
Y glasses	313		352	
Y no glasses	336	296*	308	

Tab. 1: Extracted frames for individuals X (male) and Y (female) and 4 backgrounds (sets). Images marked with * were not included in the training sets with the purpose of using some of them for testing.

Preprocessing Two training sets were constructed to represent two different domains, one with 1777 images of people wearing eyeglasses, and one with 1687 images of people wearing no eyeglasses. To test the model on unseen images, hold-out test datasets were built for each domain. Figure 2 shows some examples of the training input data. All images were additionally normalized, using a preprocessing form which brings pixel values along each dimension between -1 and 1. Since every image in the dataset is already scaled to the size of $256 \times 256 \times 3$, no further preprocessing was needed.

4.2 Results

Automatic evaluation of results is an open problem in GAN research. Unfortunately, the loss curve does not reveal much information in training GANs, and CycleGAN is no exception. Due to the nature of min-max optimization, many GAN losses do not converge, which complicates the tuning of the hyperparameters since it is difficult to apply optimization strategies such as Bayesian optimization or Grid search. For CycleGAN objective it is quite normal for both generator and discriminator losses to go up and down, as can be seen in Figure 3. One commonly used metric of quantitative evaluation for CycleGANs is the FCN (fully-convolutional

network) score. The FCN score includes standard segmentation metrics, but it is only applicable if paired ground truth data is available, which is not the case for our dataset. We present output images as qualitative results: some of the best results are shown in Figure 4 and more training set and test set results in Figure 9 in Appendix.

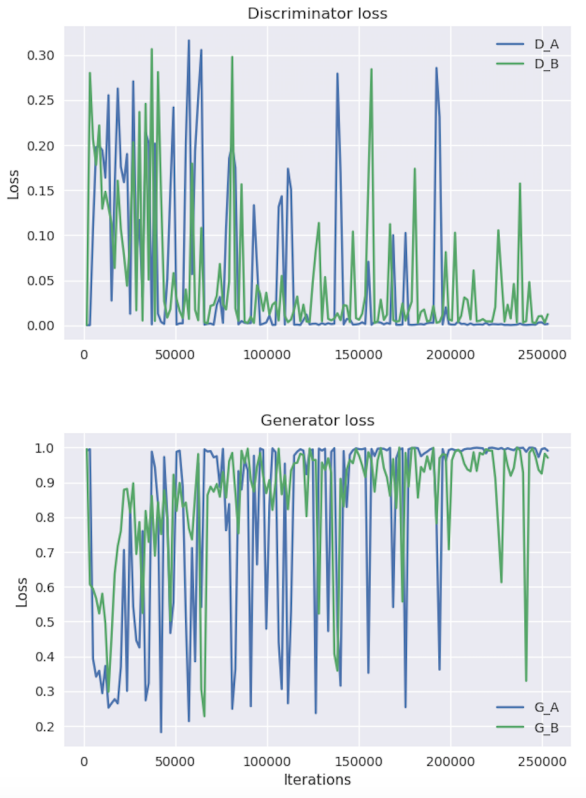


Fig. 3: Discriminator and generator loss functions. D_A and G_A denote the mapping from the glasses to no glasses. D_B and G_B denote the reverse mapping. In the implementation cycle loss is included in the generator loss.

CycleGAN experiments We noticed that the model is very sensitive to initialization. Zhu et al. also report this kind of problem. Sometimes the generators, instead of trying to generate realistic images, learn color inversion mapping so that they can collectively decrease cycle consistency loss. Once stuck in such local modes early in training, the generators are unlikely to unlearn this inversion. In this case it is suggested to restart the training, and to train the model multiple times to achieve better results. It often happens that the model succeeds in one epoch but fails in the next one, sometimes this behaviour is quite random and is not influenced by the training advances. Some typical failure cases can be found in Figure 5.

Figure 6 shows how generated image and its reconstruction are changing with number of epochs. One can notice that for the transformation from

Fig. 4: Examples of selected successful results of unsupervised image translation on the train set.

Fig. 5: Examples of typical failure cases from the train set. We can observe the color inversion in the second column.

the domain B to the domain A (faces with glasses to faces with no glasses) in the early epochs the glasses are still visible and become less and less as the training continues.

We also observe that the cycle consistency guides the training by quickly driving generators to output images similar to inputs with simple color mappings. As can be noticed from Figure 6, the generator learns a near-identity mapping as early as training epoch two.

Ablation studies We conducted ablation studies, in which we trained models for both tasks (taking off and putting on eyeglasses) without cycle and in one direction. We trained one model for the transformation from glasses domain to no glasses domain($G_a : B \rightarrow A$) and one for the reverse mapping ($G_b : A \rightarrow B$). Each model contained one discriminator and one generator, and was trained with discriminator and generator losses only. Generated pictures were not promissing and did not resemble the real distribution. Example results and loss function plots can be found in Appendix in Figures 7 and 8.

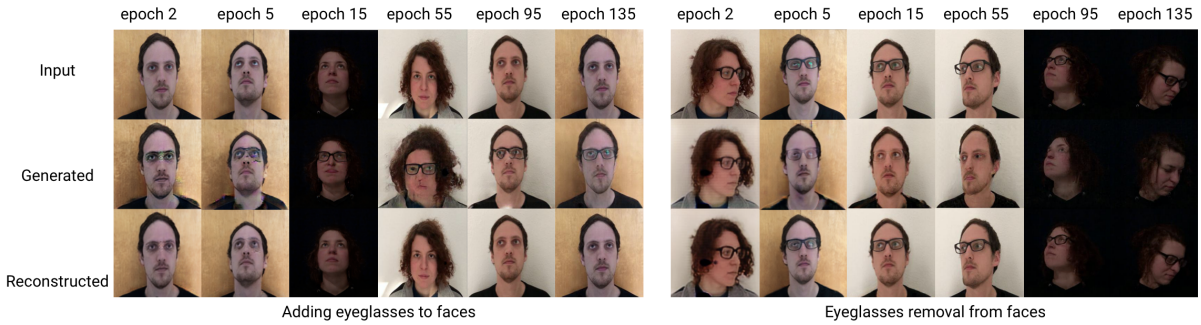


Fig. 6: The development of image generation and reconstruction during the training for both tasks.

5 Discussion and Conclusion

In this project we investigate the use of a CycleGAN framework for the task of face attribute manipulation, specifically automatic eyeglasses removal and adding eyeglasses. The ability of the CycleGAN to learn mappings between two discrete unpaired collections of images makes it appealing for face recognition domain. Experimental resuts demonstrate that the framework is not only able to learn color and texture changes, like style transfer as reported by Zhu et al. [1], but can also successfully manipulate face images while leaving most details in attribute-irrelevant areas unchanged. The great instability in training and the impact of initialization indicate that future work is necessary. Below we list some of the ideas that can be tested to improve our model.

Better Cycles is a collective name for improvements proposed by Wang and Lin [12]. Cycle consistency loss helps to stabilize the training in early stages, but it becomes problematic for generation of realistic images in later stages and enforcing cycle consistency on cycles where generated images are not realistic slows down training. Wang and Lin propose changes to cycle consistency, such as to gradually decay the weight of cycle consistency loss λ or to weight cycle consistency loss by the quality of generated images.

Wasserstein loss The cost function of a Wasserstein GAN (WGAN) [13] uses Wasserstein distance (measure of a distance between two probability distributions) that has a smoother gradient everywhere. WGAN is known to enhance training stability. Additionally, WGAN loss function does converge and hence it can serve as a measure for the image quality. The effects of adding Wasserstein loss to our CycleGAN settings could be interesting to investigate.

Identity loss Using an identity loss [14] can regularize the generator to be close to an identity mapping when presented real samples from the target domain. The generator trained with this loss could be more conservative for unknown content.

Generator architecture choices are another matter that requires further investigation. Altough in most of the applications of CycleGAN to different tasks and different domains residual networks are found to produce the strongest results, other architectures and hyperparameters capable of handling more varied and extreme transformations are an important question for future work.

References

[1] J.-Y. 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*, pp. 2223–2232, 2017.

[2] M. LIANG, Y. XUE, K. XUE, and A. YANG, "Deep convolution neural networks for automatic eyeglasses removal," *DEStech Transactions on Computer Science and Engineering*, no. aiea, 2017.

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

[4] W. Shen and R. Liu, "Learning residual images for face attribute manipulation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4030–4038, 2017.

[5] J. Guo, X. Zhu, Z. Lei, and S. Z. Li, "Face synthesis for eyeglass-robust face recognition," in *Chinese Conference on Biometric Recognition*, pp. 275–284, Springer, 2018.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

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

[9] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017.

[10] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2107–2116, 2017.

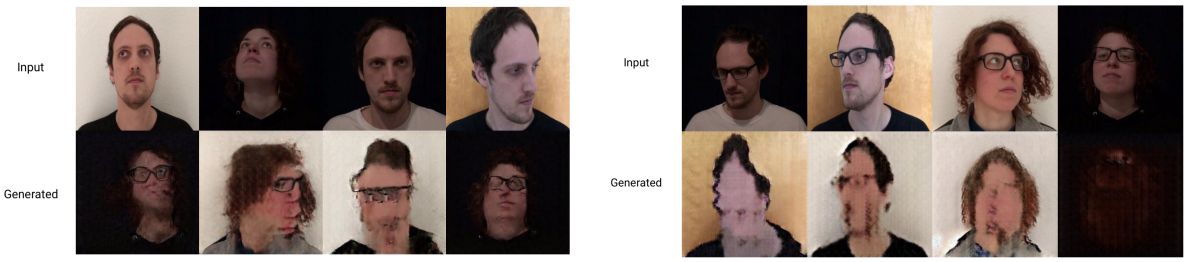
[11] S. Tripathy, J. Kannala, and E. Rahtu, "Learning image-to-image translation using paired and unpaired training samples," *arXiv preprint arXiv:1805.03189*, 2018.

[12] T. Wang and Y. Lin, "Cyclegan with better cycles,"

[13] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

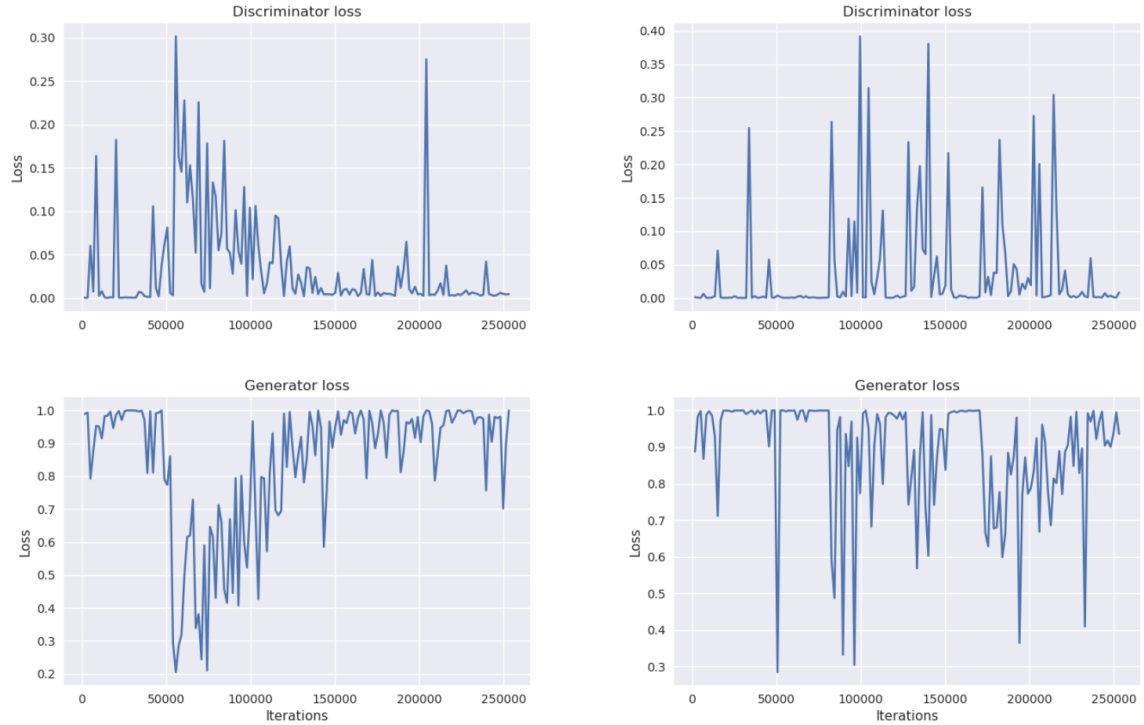
[14] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," *arXiv preprint arXiv:1611.02200*, 2016.

A Appendix



(a) No glasses domain to glasses domain translation. (b) Glasses domain to no glasses domain translation.

Fig. 7: Random results from ablation studies (one direction without cycle).



(a) No glasses domain to glasses domain translation. (b) Glasses domain to no glasses domain translation.

Fig. 8: Discriminator and generator losses from the ablation studies (one direction without cycle).

(a) Training set.

(b) Test set.

Fig. 9: Randomly selected results.

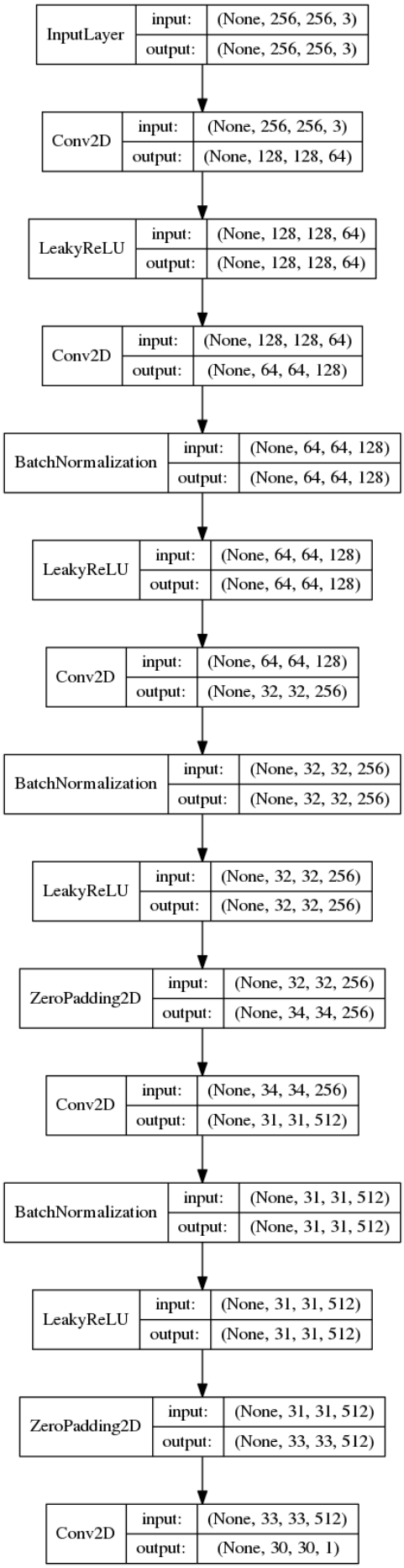


Fig. 10: Architecture of the fully-convolutional discriminator.

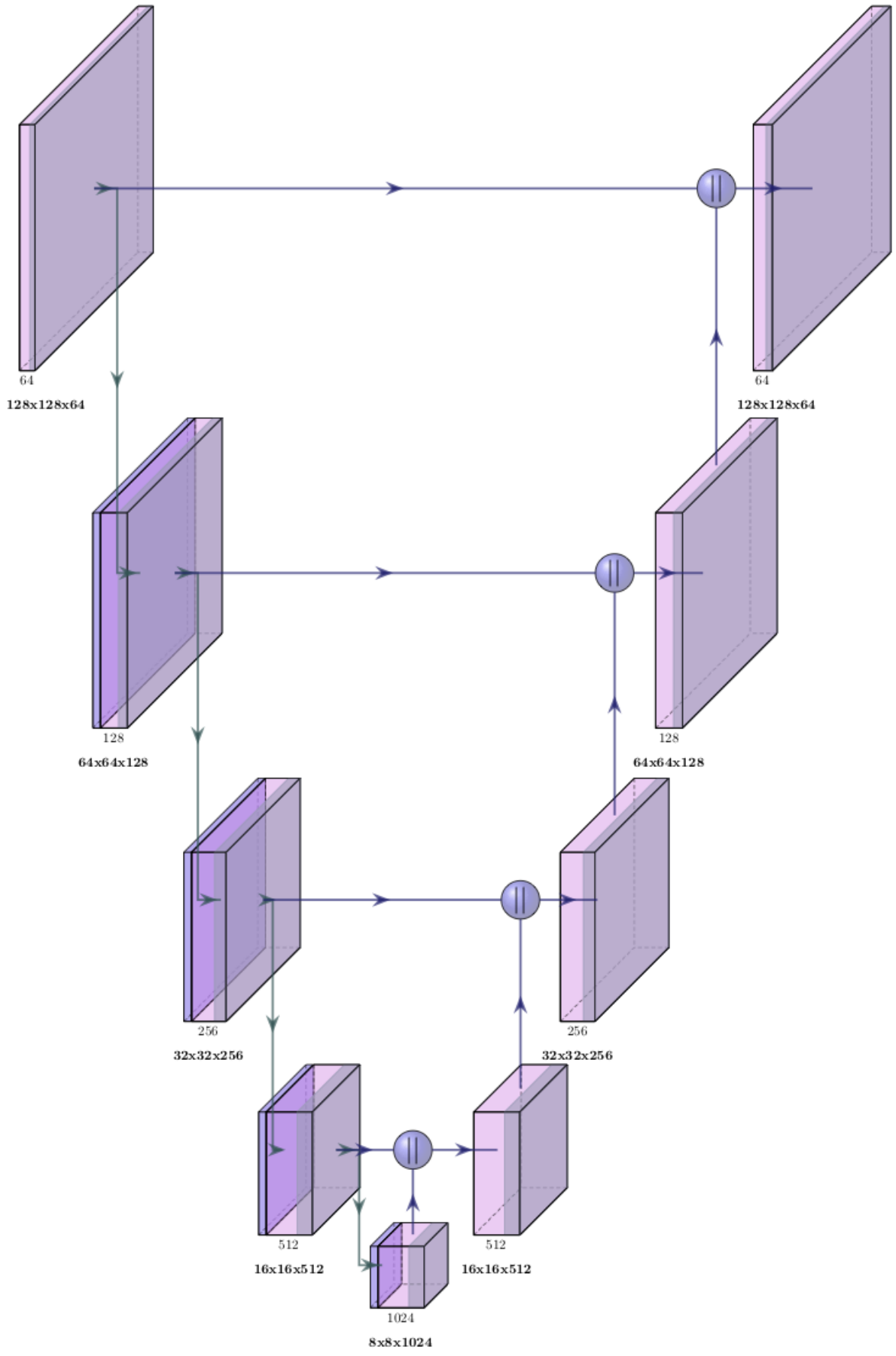


Fig. 11: Generator architecture. Input and output dimensions are $256 \times 256 \times 3$ and not shown in the figure. Each block in the contractive path contains a convolutional layer, instance normalization layer and a LeakyReLU activation layer. Each block in the expansive path contains a transposed convolutional layer, a cropping layer and a ReLU activation layer. The last layer has a tanh activation. The circles denote skip connections between contractive and expansive paths.