# Project AI: Face Image Synthesis conditioned on Target Landmark Appearance

**Elias Kassapis** (12409782)[1] **Stijn Verdenius** (10470654)[1] **Klaus Ondrag** (12265306)[1]

Source code - https://github.com/StijnVerdenius/DeepFaceImageSynthesis

## 1. Introduction

We consider the problem of generating photorealistic, personalized head models, that capture mimics and speech expressions of a particular individual. Specifically, we focus on the task of face reenactment; that is, transferring a person's (source) facial expression movements to another person's (target) face. Having a wide range of applications, including virtual reality, film production, or video conferencing, various methods have been developed for the task, many of these utilizing Generative Adversarial Networks (GANs), and yielding impressive results.

cGANs consist of a generator and a discriminator network, and learn a mapping between input and output images. Crucially, they also learn a loss function that reins the training of this mapping (Isola et al., 2017). In vanilla GANs, the loss function is trained by using an adversarial loss term that encapsulates a minimax game, where the trainable discriminator learns to output a score of "realism" for input images, and the generator learns to generate images that appear realistic in order to "fool" the discriminator (Goodfellow et al., 2014). Therefore these networks learn a loss function that adapts to the data, ensuring that the objective function is tuned to be appropriate for realizing our goal.

In this setting, various conditioning criteria have been experimented with when using cGANs, such as illumination, face expression, eye gaze, and head position (Kim et al., 2018). However, more recent approaches were able to capture both changes in expression and pose by using the source image as input, and only conditioning on target landmarks (Sanchez & Valstar, 2018; Zakharov et al., 2019). Various loss term extensions have been developed to facilitate task-related optimization, which fall mainly into two categories: pixel space, and feature space loss terms. Pixel space loss terms minimize the distance between each pixel of a pair of the ground truth target image and the generated image, and feature space loss terms minimize the difference between extracted features of the said pair.

Even though GANs can generate high quality, realistic faces, identity preservation is an issue. Identity information such as hair colour and eye colour are not always preserved, as the networks tend to prioritize generating plausible faces over maintaining the identity of the subject (Li & Luo, 2017; Sanchez & Valstar, 2018). This is the case especially when we use extreme or unseen poses, where the generated images can appear completely unrealistic, and rendered with artifacts.

In the current work our aim was to investigate how these loss terms affect identity preservation in face reenactment using cGANs. We hypothesized that feature difference minimization loss terms are more important for identity preservation that pixel minimization loss terms. Therefore, we employed an ablation approach to dissect the effect of the different elements in the loss function on identity preservation by qualitatively evaluating their performance on face reenactment tasks. We based our implementation on the pipeline developed by (Isola et al., 2017), and used a 6-term loss function adapted from (Sanchez & Valstar, 2018), which consists of 4 different pixel space terms, and 2 feature space terms. We grouped our loss terms into groups feature difference minimization terms and pixel difference minimization terms, and ablated a different group for each experiment.

We found that both pixel and feature space terms are important for identity preservation, consistent with previous reports, however we could not show which is more important in order to accept or reject our hypothesis. We concluded that it is the synergistic interaction between the feature space terms and pixel space terms that promotes identity preservation, as these appear to have complementary merits. Feature space terms appear to facilitate target conditioning, and pixel space terms appear to encourage the generator to regress to identity mapping.

## 2. Background

In this work we implemented a cGAN-based model used to generate a set of person-specific frames driven by a set of target landmarks, allowing us to manipulate a face in a realistic fashion. Our goal was to investigate the effect of different loss terms used in the literature for GAN-based approaches for face reenactment in order to dissect which terms govern identity preservation in this process.

## Models

This section describes the architecture details of the investigated models. We adopted the pipeline proposed in the image-to-image paper (Isola et al., 2017), as this model has showed good results for face synthesis (Sanchez & Valstar, 2018). The models are described below.

### U-Net Generator

The U-net generator is an architecture first proposed in (Ronneberger et al., 2015). It consists of several down-sampling convolution blocks (in our case 4), then one resnet block in the bottleneck and consequently up-sampling convolution blocks. The unique aspect of this generator is that it contains skip-connections from the down-sampling flow to the up-sampling flow at equal level from the input/output layer respectively. This helps maintaining identity and has been shown to increase visual quality (Isola et al., 2017).

### PatchGAN Discriminator

The PatchGAN discriminator is a discriminator that, instead of the traditional GAN-based discriminator network that looks at the complete data sample, looks at patches of the input sample and determines if these are realistic patches on average, instead of judging the entire sample in one go. This increases sharpness in pictures at patch level (Isola et al., 2017). Our type of PatchGan was first introduced in (Li & Wand, 2016) for that reason. In our setup 3 down-sampling convolutions are used and thereafter the mean operation. In addition, we use 32 number of hidden channels per down-sampling and a lower learning rate for the discriminator. Moreover, the input is enriched with target landmarks to ensure conditioning.

## 3. Methods

## Task and Data

We used the 300VW dataset (Shen et al., 2015), which is a collection of 114 videos which consist of over 200,000 images. With every image, $n_{landmarks}$ face landmarks are stored. The videos typically feature a single person facing the camera and talking. We processed the dataset by extracting the frames, taking a square around the center of the landmarks for each frame and saving the cutout. We also offset the positions of the landmarks and saved the landmarks in an array for each video. This was done because the loading of a lot of small files takes much longer and all landmarks can easily be held in memory. For passing the landmarks into the models, they are first converted from $(n_{landmarks}, 2)$, where 2 represents the x and y coordinate, to $(imsize, imsize, n_{landmarks})$. This was done so that the models can distinguish between the different landmarks.

In order to enable the models to learn better, we applied a Gaussian in the area around the landmark coordinate instead of having one value being 1 and the rest 0. For calculating the losses, we required three image-landmarks-pairs for a single forward and backward pass. These always belong to the same video. As for augmentations, we implement random horizontal flips. These are applied either to all three samples images and landmarks or not at all. We also experimented with random cropping but this degraded the performance.

## Training

The dataset was split according to the official guideline into training (50 videos) and three test sets (31, 19 and 14 videos). Regarding the number of landmarks, $n_{landmarks} = 68$. The $imsize$ is 64. The Gaussian is applied over 7 pixels with a mean of 0 and a variance of $1/3$.

We used Adam optimizer with learning rate = 2e-4 to train the generator, and learning rate = 5e-5 to train the discriminator, for 6 epochs on one NVIDIA GTX 1080 GPU. Over the 6 epochs the model is trained on a total of 57.2k samples of 3 video-specific image-landmark pairs, which is 9.5k of said samples per epoch. Our loss function for the generator $\mathcal{L}_G$ is formulated by 6 terms, defined by

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{pix}\mathcal{L}_{pix} + \lambda_{self}\mathcal{L}_{self} + \lambda_{triple}\mathcal{L}_{triple} + \lambda_{pp}\mathcal{L}_{pp} + \lambda_{id}\mathcal{L}_{id}$$

where $\lambda_{pix} = 10$, $\lambda_{self} = 100$, $\lambda_{\text{triple}} = 100$, $\lambda_{pp} = 10$ and $\lambda_{id} = 1$. $\mathcal{L}_{pix}$, $\mathcal{L}_{self}$ and $\mathcal{L}_{triple}$ are pixel distance minimizing loss terms, and $\mathcal{L}_{id}$ and $\mathcal{L}_{pp}$ are feature distance minimizing loss terms. These are analyzed below.

#### Adversarial Loss

The Adversarial loss term ($\mathcal{L}_{adv}$) drives the distribution of the generated images to be the same as the distribution of the training images. We used a binary cross-entropy (BCE) loss for the discriminator $\mathcal{L}_D$ given by

$$\mathcal{L}_D = -\frac{1}{N}\sum_{j=1}^{N}\left(y_j \log\left(\hat{x}_j\right) + \left(1 - y_j\right)\log\left(1 - \hat{x}_j\right)\right)$$

where N is the number of generated samples in the current batch, $y_j$ is the ground truth label, and $\hat{x}_j = D(\hat{I}_j)$ is the output of the discriminator for the generated images, $\hat{I}$. Now the loss for the generator $\mathcal{L}_{adv}$ is defined by

$$\mathcal{L}_{adv} = -\mathbb{E}_{\hat{I}\sim B}[\log(\hat{x})]$$

where $B$ is the set of the generated images, $\hat{I}$, in the current batch.

### Pixel Loss

The Pixel loss term ($\mathcal{L}_{pix}$) conditions the expression of the subject in the generated target image, $\hat{I}_t$ to match the desired one $I_t$. This is given by

$$\mathcal{L}_{pix} = \|G(I_s; \mathbf{e}_t) - I_t\|_2^2$$

where $G(\cdot)$ is the generator, $I_s$ is the source image, $\mathbf{e}_t$ are the landmarks of the target image $I_t$.

### Self-Consistency Loss

The Self-Consistency loss term ($\mathcal{L}_{self}$) is a term widely used to preserve identity (Sanchez & Valstar, 2018). This is done by first generating an image based on target landmarks, then take the source landmarks and revert the image back to the original and finally taking a squared L1-loss distance. This is defined by

$$\mathcal{L}_{\text{self}} = \|G(G(I_s; \mathbf{e}_t); \mathbf{e}_s) - I\|^2$$

### Triple-Consistency Loss

The Triple-Consistency loss term ($\mathcal{L}_{triple}$). defined by (Sanchez & Valstar, 2018), ensures consistency of progressive generation of images. Practically, we compare a direct generation of the target with a generation of the target conditioned on another generated image using 'in-between' landmarks. Again a squared L1-loss distance is used. This is given by

$$\mathcal{L}_{\text{triple}} = \left\|G\left(\hat{I}; \mathbf{e}_t\right) - G(I; \mathbf{e}_t)\right\|^2$$

### Perceptual Loss

The Perceptual loss term ($\mathcal{L}_{pp}$) encourages the generator to reproduce subtle details expressed in the training images by enforcing the features of the two to be similar. We use the perceptual loss defined by (Sanchez & Valstar, 2018), where the perceptual loss is decomposed into a feature reconstruction loss, and a style reconstruction loss term. The real images and generated images are forwarded through a VGG-19 network, and the difference between the activations $\Phi^l_{VGG}$ at layers $l = relu1\_2,\ relu2\_2,\ relu3\_3,\ relu4\_3$ is computed using the L1-norm for the feature reconstruction loss, whereas the style reconstruction is evaluated by computing the Forbenius norm of the difference between the Gram matrices of the activations at the $relu3\_3$ layer of the VGG-19 network. The perceptual loss term is then defined as

$$\mathcal{L}_{pp} = \sum_l \left\|\Phi^l_{VGG}(I) - \Phi^l_{VGG}(\hat{I})\right\| + \left\|\Gamma\left(\Phi^{\text{relu3}-3}_{VGG}(I)\right) - \Gamma\left(\Phi^{\text{relu3}-3}_{VGG}(\hat{I})\right)\right\|_F$$

where $\Gamma$ are the Gram matrices.

### Identity Loss

The Identity loss term ($\mathcal{L}_{id}$) is used to promote the preservation of identity information. We use the Identity loss defined by (Shiri et al., 2018), defined as the L1 distance between the activations $\Phi^l_{VGG}$ at layer $l = relu2\_3$ of the VGG-19. This is given by

$$\mathcal{L}_{id} = \left\|\Phi^l_{VGG}(I) - \Phi^l_{VGG}(\hat{I})\right\|$$

## 4. Experiments

## Grouping loss terms

We grouped loss term extensions in three groups, the first two groups only consisting pixel-space terms, and the last group only feature-space terms. These are: (1) Pixel Loss, (2) Consistency Losses, (3) Perceptual and Id Losses. We grouped the terms in this fashion so that we have a "control" ablation for the pixel loss terms, referring to the first group, to assess any implicit contribution to identity preservation, and we had a term in each of the other groups used to explicitly preserve identity, referring to the self-consistency loss, and id loss.

## Ablation study

The main experiment of this paper is an ablation study of the different groups of loss term extensions. This lead to 4 runs: (1) full loss function, (2) pixel loss ablated, (3) consistency losses ablated, (4) perceptual and id loss ablated. We thereby directly evaluate whether pixel-space based loss terms exert a stronger influence than feature-space based terms on the preservation of identity.

## Evaluation

To evaluate the performance of our model using each version of the loss function, we qualitatively compared the generated images focusing on degree of conditioning to target landmarks and preservation of identity. We also generate heatmaps by taking the absolute difference between the generated image and the target image. Afterwards, we average over the color channels. We visualize the result with the viridis color palette, where blue represents 0 and yellow values above 96. Although the errors could range from 0 to 255, we found that this setup allows for a better visual inspection because the images are much alike.

# 5. Results and Analysis

## Ablation study

Results of the ablation study are displayed in Appendix A. It demonstrates qualitative results of the 4 ablation experiments graphically. As mentioned before, we compare results on the basis of our own judgment on the degree of identity preservation and conditioning-ability on the generated images.

There is a number of things that stand out. Initially, it appears that the generated images of the ablated identity and perceptual loss run (Figure 1c) have a good identity preservation, yet little-to-no conditioning. We expected that it would be detrimental to identity preservation, but would not significantly affect conditioning, as the pixel and triple consistency losses have been shown to promote target conditioning in previous work (Isola et al., 2017; Sanchez & Valstar, 2018; Zakharov et al., 2019). However, our results indicate that our generator has collapsed into learning identity mapping between input and output. Taking into account the magnitudes of the loss terms in the final epoch, as shown in Figure 3c, we can see that the self-consistency has a higher magnitude than the rest of the terms, which indicates that it dominates over them. Our interpretation is that since the self-consistency term minimizes the difference between the source image and generated image conditioned on source landmarks, this encourages the generator to learn identity mapping, especially considering the residual connections in the generator, as it is an easier task than learning progressive generation. Nevertheless, contrary to our expectations, using the current configuration of loss term weights, feature-space loss terms do not appear to promote identity preservation but instead appear to be the main force of conditioning.

Our results from ablation of the consistency loss terms (self and triple) points out that these are responsible for a great deal of identity preservation. In Figure 1a, we observe that without them the generated image has glasses and a lighter illumination as compared to the other figures in Appendix A, in which the consistency losses were present during training. We argue that the glasses appear because these appear in multiple specimens in our dataset, and they always have a similar shape and appear at the same location on faces (over the eyes). Therefore their emergence indicate that realistic face synthesis is prioritized over identity preservation. Again contrary to our hypothesis, given our results, pixel-space terms appear to be more important than feature-space terms for identity preservation.

Considering the ablation of the pixel loss we see that a lot more irregular behaviour. It appears that both the conditioning is affected, and the identity preservation. Looking at Figure 1b, we can once again see traces of glasses, therefore, we can deduce that pixel loss also implicitly affects identity preservation. Again this illustrates that pixel-space losses are indeed quite important for identity preservation.

Finally, the model trained on the total loss function (displayed in Figure 1d) appears to both do conditioning and identity preservation to a reasonable extent, and clearly outperforms the models trained with the different ablations.

## Heatmaps

For a second evaluation heatmaps of error were made to illustrate the differences between the ground truth target image and generated image. These can be observed in Appendix B.

# 6. Conclusion

Given our results we cannot accept our hypothesis that feature-space loss terms are more important for identity preservation. However, given our observation that when ablating the feature-space loss term group, self-consistency overpowers the rest of the pixel terms leading to the generator to collapse to identity mapping, our results do not provide a fair comparison. We expect that if we completely remove self-consistency or decrease it's corresponding weight, we would observe conditioning, and a greater effect on identity preservation. Therefore, we cannot reject our hypothesis either. We conclude that there is a synergy between the feature space terms and pixel space terms that enhances identity preservation, as these appear to complement each other.

However, in this research a number of setbacks and limitations were experienced. Initially, the main limitations of our study is that we did not ablate individual terms, and we did not include any quantitative evaluation of our results. This was mainly due to limited time and computational resources. This limits our conclusive power as it is hard to compare results to other resources. As a consequence, real hard statistical analysis was also inconceivable. An interesting extension of our study would be to research at individual loss functions in the same experimental setting.

Finally, a substantial limitation of our work is the fact that we had access to target images. In most face synthesis situations, only the target landmarks are available, not the target image itself as well. This would make some of the loss functions used obsolete as they rely on the availability of target images. Think for example of attempting to synthesizing videos of historical or fictional characters such as Einstein or characters in video games. In these cases target ground truth images are absent.

## 7. Privacy and ethics

The subject of generating human images has sparked discussions concerning privacy and ethics. Even before the digital age, photo manipulation was possible. It became more accessible in the digital age with tools like Photoshop and video editing software. The work in this project is progress in the same direction by making the creation of images depicting humans easier. We, the authors agree that these tools can cause harm in the wrong hands.

Recently, they might have also been used in a political context to influence public opinion. This potentially big impact on society needs to be addressed, both morally and legally. Because of this, we wanted to explore the current state of the art. The shortcomings and limitations of this and similar work can help us develop tools to stop fake videos. For example, since our model was processing frame by frame, it lacked temporal consistency. Metrics that measure this might be able to flag videos which are manipulated.

[1]Supervisor: Minh Ngo and Sezer Karaoglu.
Elias Kassapis <EliasKassapis@student.uva.nl>
Stijn Verdenius <stijn@verdenius.com>
Klaus Ondrag <klaus.ondrag@student.uva.nl>.

*Project AI 2019*, MSc Artificial Intelligence, Univ. of Amsterdam

## References

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf. 1

Isola, P., Zhu, J.-Y., Zhou, T., and 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. 1, 2, 4

Kim, H., Carrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C., Zollhöfer, M., and Theobalt, C. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):163, 2018. 1

Li, C. and Wand, M. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pp. 702–716. Springer, 2016. 2

Li, Z. and Luo, Y. Generate identity-preserving faces by generative adversarial networks. *arXiv preprint arXiv:1706.03227*, 2017. 1

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer , 2015. 2

Sanchez, E. and Valstar, M. Triple consistency loss for pairing distributions in gan-based face synthesis. *arXiv preprint arXiv:1811.03492*, 2018. 1, 2, 3, 4

Shen, J., Zafeiriou, S., Chrysos, G. G., Kossaifi, J., Tzimiropoulos, G., and Pantic, M. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 50–58, 2015. 2

Shiri, F., Porikli, F., Hartley, R., and Koniusz, P. Identity-preserving face recovery from portraits. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 102–111. IEEE, 2018. 3

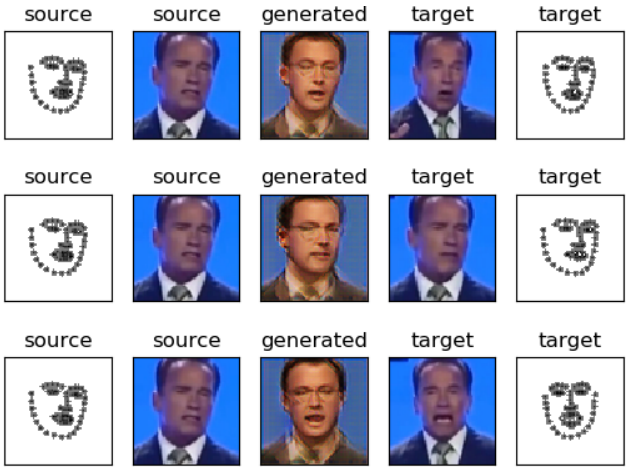
Zakharov, E., Shysheya, A., Burkov, E., and Lempitsky, V. Few-shot adversarial learning of realistic neural talking head models. *arXiv preprint arXiv:1905.08233*, 2019. 1, 4

## Appendix A: Ablation study results

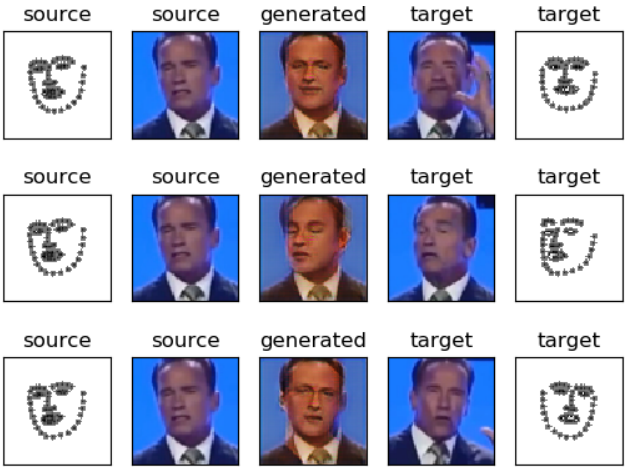
In this section result examples are displayed concerning the ablation study. Each of the 4 runs has its own figure (Figures 1d, 1c, 1a & 1b). Each figure contains 3 samples from the model trained with the in the figure mentioned loss functions, which are displayed in the rows. Each has source picture and landmarks (first two columns), target picture and landmarks (last two columns), and the generated picture in the middle column. For all models is tested on the same test video (of Arnold Schwarzenegger). Additionally, in figure 2 generations of multiple specimens are displayed with the four trained ablation models. Added is their heatmap error figures to demonstrate the error-sensitive areas in generation.

*Figure 1.* Ablation study results

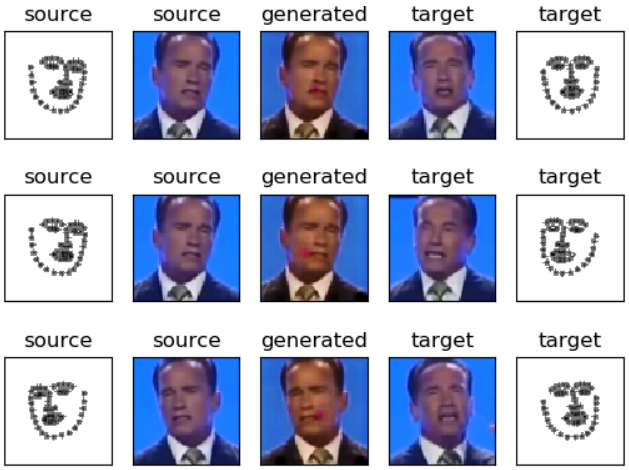
(a) Trained with ablated consistency losses.



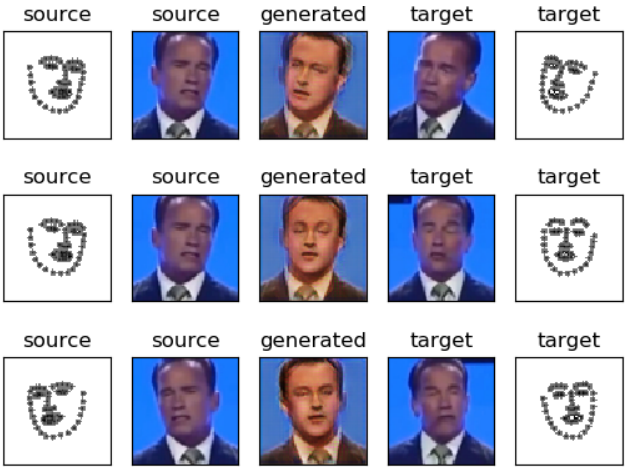

(b) Trained with ablated pixel loss.




(c) Trained with ablated perceptual and identity losses.




(d) Trained with full loss.



## Appendix B: Heatmaps

Figure 2 shows generated images from different models. The first column displays the landmarks of the input image. This is not passed into the network. The second and third column are the image to condition on and the target landmarks. Both are the input to the network. In the fourth column, "Target Image", the ground truth image is shown, which displays the same person as in the input image and has the landmarks of the input landmarks. The next four columns show the generated images of the model trained with different loss functions. The last four columns show the absolute pixel-wise error averaged over the color channel, on a color scale where 0 is represented in blue and values above 96 in yellow.

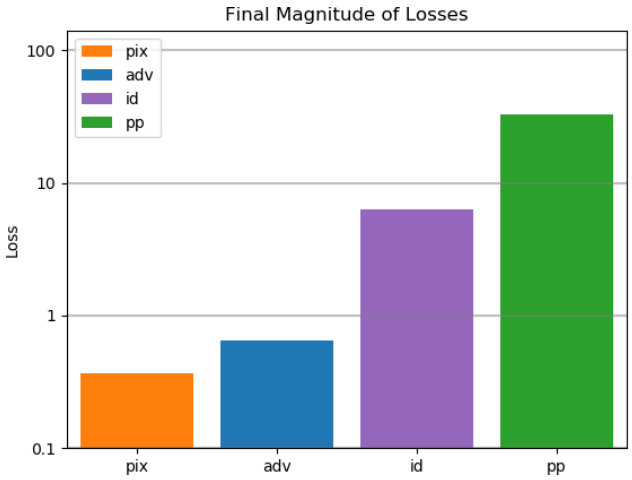

*Figure 2.* Heatmaps with their respective generations for a number of sample source-target-pairs.

## Appendix C: Other

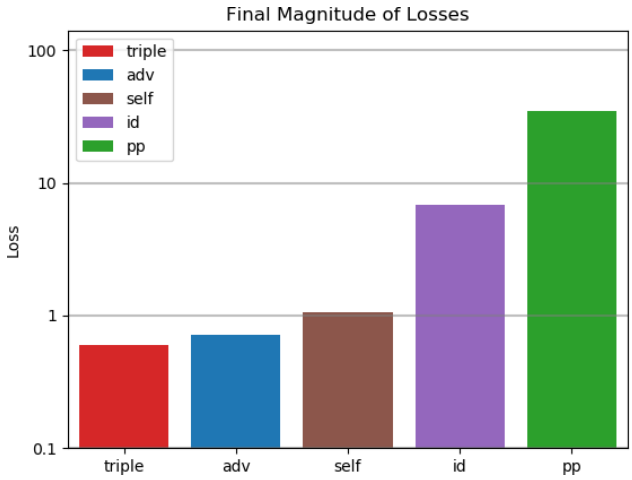
In this section results are displayed concerning the training progress. The results of the final loss magnitude in Figures 3c, 3b, 3a & 3d. Training progress over time is displayed in Figures 4a & 4b.

*Figure 3.* Bar plot of magnitudes of individual loss functions at the end of training.

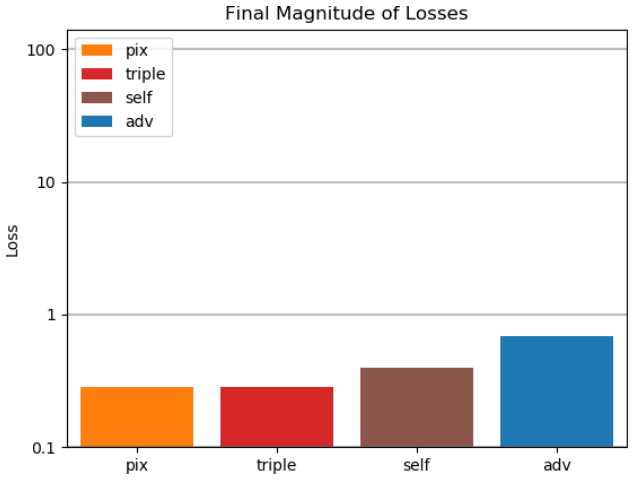
(a) Trained with ablated consistency losses.




(b) Trained with ablated pixel loss.




(c) Trained with ablated perceptual and identity losses.




(d) Trained with full loss.

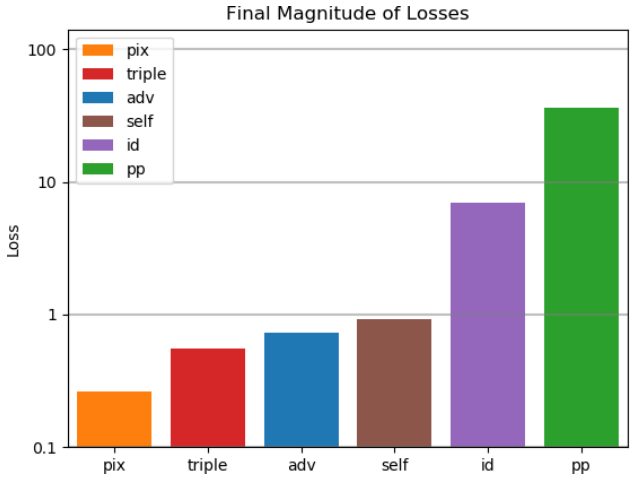
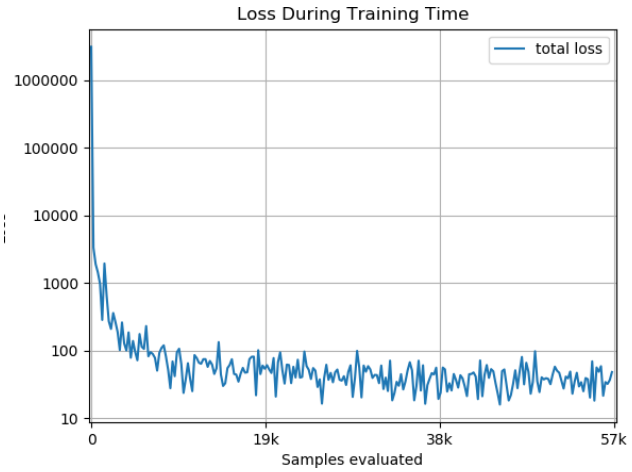


*Figure 4.* Losses over training-time during the run with all loss functions.

(a) Total loss function




(b) Individual loss functions