More Mori: Face Swap with Limited Data

Alex H. Nguyen
Simon Fraser University
Burnaby, BC, Canada
ahn5@sfu.ca

Sean Beaudoin
Simon Fraser University
Burnaby, BC, Canada
sbeaudoi@sfu.ca

Shuman Peng
Simon Fraser University
Burnaby, BC, Canada
shumanp@sfu.ca

1 Introduction

There are plenty of fish in the sea, but only one Mori. *MoreMori* aims to transfer the look and features of Dr. Mori onto any person using readily available data from online sources. We target facial features for transfer and add the additional constraint of computing the results in realtime to support a live demo at reasonable framerates.

The goal of a face transfer is to generate an image where the original face is transformed in such a way to appear plausible that it is actually someone elses face. Further more, a transfer should preserve some of the original face features such as expression and pose. The hair, neck, ears and torso of the person should remain unchanged. Although many methods exist for computing a transfer, we aim to explore the capabilities of generative adversarial networks (GAN) for this task. We attempt to use methods which generalize to other object transfigurations instead of specializing to facial transformations.

CycleGAN[2] is a generative adversarial network which adds the cycle consistency constraint which measures the ability of the network to reverse transformations.

A *CycleGAN* has the following functions[1]:

- Two domains X and Y
- Two mappings $G : X \to Y$ and $F : Y \to X$
- Two discriminators:
  - D_x : distinguishes between $\{x\}$ and $\{F(y)\}$
  - D_y : distinguishes between $\{y\}$ and $\{G(x)\}$
- Two cycle consistencies:
  - Forward: $x \to G(x) \to F(G(x)) \approx x$
  - Backward: $y \to F(y) \to G(F(y)) \approx y$

For a mapping, the loss function is

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{\text{data}}(y)}[\log D_y(y)] \\ + \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(1 - D_y(G(x)))]\end{aligned}$$

For cycle consistency, the loss function is

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

Together, the loss function for *CycleGAN* is

$$\begin{aligned}\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) \\ + \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) \\ + \lambda \mathcal{L}_{\text{cyc}}(G, F)\end{aligned}$$

[1] Equations Zhu et al[2]

A variety of previous work exists regarding the use of *CycleGAN* for style transfers. *CycleGAN* has wide-ranging applications, and is not specific to human faces. DeepFake[2] builds on *CycleGAN* by mapping face-specific features to assist in alignment of features during transfer. Attention GANs[3][4] generate attention masks used by the generator and discriminator of the GAN, which aids in applying transformations only within the salient region.

Existing facial feature transfer models translate one person's face into another utilize features specific to faces to generate alignments. The goal of *MoreMori* differs in this respect as it is a many to one translation and can be generalized to perform translation of objects other than human faces.

2 Approach

We used a data-centric approach, leaving the networks of *CycleGAN* unmodified. This allowed us to explore the effects of data inputs on training and testing results with a known-good network.

2.1 Data Collection

In order to collect representative data, a pipeline was configured to collect data from popular online sources in the form of video, which more closely matches the intended input than still images. An alternate entry point allows the user use any video as input. A sampling rate was determined by the length of the videos and volume of data available for each subject to build a diverse dataset. A cropped and resized image was collected from each video frame sampled using a pre-trained deep learning face detector to obtain the region of interest.

Initial training efforts using the plain Mori data from the two available online videos (approximately 30 seconds[3] and 50 minutes[4] of footage) resulted in transfer of the background region as well as poor retention of the original facial pose. Models from early epochs were unable to generate a recognizable face when used with unseen test images. As a result, images of an alternate subject Alex were collected to investigate the performance of CycleGAN with a more diverse dataset. With full control over the data collection, it was possible to acquire images with a variety of background and lighting conditions. Models built on this data referred to as *World vs Alex*[5] showed improved performance, however properties of the Alex data's backgrounds, such as the average colour would persist in the transform images.

2.2 Segmentation

Previous works from Chen et al[3] and Mejjati et al[4], employ a separate network to generate masks for the input images. In Chen et al[3], the masks are used to control combination of the input and generated images. Mejjati et al[4] differs from Chen et al[3] in that masks are used directly in the generator and discriminator loss functions in addition to controlling the combination of input and generated images. The use of attention in the generator and discriminator loss functions allows the background to become unconstrained with no effect on loss. Although these methods improve on CycleGAN, they both introduce an additional network required for model evaluation.

To mimic attention masking, we first explored generating a mask to segment the training input images. A popular fully convolutional network[1] was used generate the masks identifying face regions in the image. Initial results using the segmented faces on black backgrounds showed improvement over models trained with their original backgrounds. Segmented images were evaluated on models trained with segmented images. Unfortunately it was discovered that accurate segmenters incur significant runtime cost, so while this method improves face transfer results, it is ultimately too slow for real time use on the target hardware. Evaluating unsegmented images with the models trained on segmented inputs produced dark outputs, making this model unsuitable for use without a segmenter to preprocess inputs. The results of this model however did expose the potential of *CycleGAN* to learn segmentation of images which we explore in the final solution.

[2] https://github.com/deepfakes/faceswap

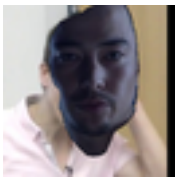
[3] https://youtu.be/yt2VQKOqJCI

[4] https://youtu.be/D-o6c9ti9RE

[5] *World vs* ***Person*** refers to the model trained using a dataset that contains images of random people and a dataset that contains images of ***Person***.

2.3 Background Swap (BGSwap)

(a) Dr. Mori's face placed on a person



(b) A person's face placed on Dr. Mori

(c) Dr. Mori's face placed on a person

The final solution substitutes black backgrounds previously paired with segmented faces for random backgrounds. The background image is an unaltered image from the union of datasets A and B onto which a segmented image is placed. Placement of the segmented face images does not necessarily correspond to the background face, thus the original background face may still be visible in the composite image. Pairing of the segmented face with a randomly selected background was performed in the data loader to allow image augmentation to be applied to the resulting composite image.

By diversifying the backgrounds, the GAN discriminator can no longer use background information to identify fakes. Although this should leave the generator free to produce random backgrounds, cyclic losses keep the backgrounds constrained. An identity mapping of the real background to the fake background minimizes cyclic losses without incurring additional discriminator losses. Our experiments indicate that face segmentation is learned with this method. By building the segmenter into the GAN generator network, one network can be eliminated from the evaluation process. Model evaluation is performed on unsegmented input images.

2.4 Three Stage Transfer Learning

One major difference between the performance of *World vs Alex* and *World vs Mori* models is the robustness to varied facial expressions and poses. To combine the robustness of the *World vs Alex* model with the properties of Dr. Mori's face, a common method *transfer learning*[10] was utilized. Initially, a *World vs Mori* model was trained using *World vs Alex* as a base. Various base models each trained for a different number of epochs were used as a starting point, however it was discovered that Alex and Mori were too different for the transfer to be effective.

A lookalike mode referred to as MoreToki[6] was selected based on the availability of varied data and facial similarity to Dr. Mori. Although the transfer was more successful with this lookalike, the resulting models were less robust to unseen facial expressions and poses than the Alex models.

A three stage transfer learning process was devised to incorporate a greater variety of facial expressions and poses, while retaining the look of Mori in the final model. A base model was trained using the union of Alex, MoreToki, and Mori data. Alex data was then removed, and the model trained further with the MoreToki and Mori data. Finally, the MoreToki data was removed leaving only Mori data, and further trained to fine tune the model.

2.5 Model Evaluation

The evaluation pipeline consists of face detection, resizing, *CycleGAN* model evaluation, and placement onto the original image. Multiple faces may be processed within a single image and overlaid simultaneously for real time display. The OpenCV *Haar Cascades*[5] face detection method was selected to minimize runtime cost. Minimization of runtime cost is crucial to allow multiple faces to be detected, evaluated and positioned on the image without perceivable delay.

[6] https://youtu.be/v63-AnWSZxQ

3 Experiments

The *World* dataset has 7000 images of different people's faces.

3.1 Base Mori

(a) Mori dataset 1 (200 images) (b) Mori dataset 2 (600 images)

(c) Turning a person into Dr. Mori

We trained *World vs Mori* without any preprocessing. The lack of diverse backgrounds causes the model to overfit to the background. The lack of diverse facial poses also results in distortion.

3.2 Base Alex

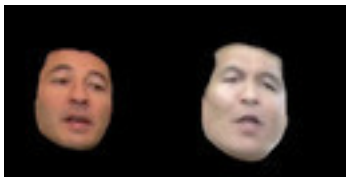
(a) Images of Alex (800 images)

(b) Turning a person into Alex

We trained *World vs Alex* without any preprocessing. Even with diverse data, *CycleGAN* changes the background [3]. The face is not distorted, but the facial features are not strongly transferred over.

3.3 Segmented Mori

(a) Segmented image of Dr. Mori

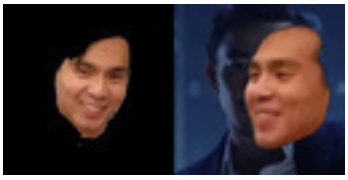


(b) Turning a person into Dr. Mori

As a sanity check, we trained the model using the simplest data possible. We trained *World vs Mori* with segmented faces. The model transfers recognizable features of Dr. Mori without background modification.

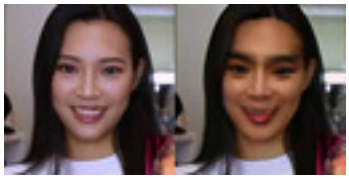
This model is not practical for real-time face swapping because segmenting images will add overhead to the runtime. The segmentation processes one face every 0.0758 seconds. This translates to 13.19 frames per second for one face in an image, and 6.60 frames per second for two faces in an image. As we want to run *MoreMori* in realtime, segmentation during evaluation is not possible.

3.4 Background Swapped Alex (BGSwaped Alex)



(a) BGSwap on Alex

(b) Turning a BGSwaped Mori to Alex

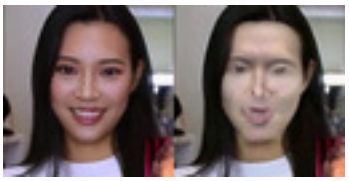


(c) Turning a person to Alex

World vs Alex with BGSwapped images has promising results. The background is unaffected and facial features are cleanly transferred over.

3.5 Background Swap Mori (BGSwapped Mori)

(a) Turning a BGSwaped person to Dr. Mori



(b) Turning a person to Dr. Mori

(c) Turning a person to Dr. Mori

We trained *World vs Mori* with BGSwapped images. The background is unaffected, but the faces are distorted. This is likely due to the lack of diverse angles and lighting for images of Dr. Mori.

3.6 Three Stage Transfer Learning

(a) MoreToki (2000 images)

(b) Turning a person to Dr. Mori

(c) Turning a person to Dr. Mori

Transfer learning solves problems with *3.5 BGSwapped Mori*. Facial features transfer over without significant background modifications. Some images such as 3.6(b) are still distorted. Amazingly, the model manages to keep the food in the image. It is a large improvement over *3.1 Base Mori*.

3.6.1 Base Phase

World vs Mori, MoreToki, and Alex

3.6.2 Transfer Learning 1

World vs Mori and MoreToki

3.6.3 Transfer Learning 2

World vs Mori

4 Comparisons

We compare our method to publicly available projects online. Deepfakes has a dataset of president Donald Trump and actor Nicolas Cage we will compare with.

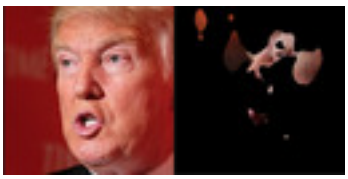
4.1 deepfakes/faceswap [7] [8]

(a) Image of Nicolas Cage

(b) Donald Trump turned into Nicolas Cage

(c) Hilary Clinton as Donald Trump (from youtube)

4.2 End-to-end, automatic face swapping pipeline [9]



(a) Segmented Image of Donald Trump

(b) Segmented Image of Nicolas Cage

We try the method presented by Nirkin et al[1], however we experienced poor segmentation performance on the Donald Trump and Nicolas Cage dataset. We did not further pursue this pipeline due to time constraints.

4.3 Our Method

(a) Segmentation fix

(b) Donald Trump turned into Nicolas Cage

(c) Donald Trump turned into Nicolas Cage

We test our method using BGSwap without transfer learning. Our method uses the segmentor from Nirkin et al[1], however modifications were required to achieve reasonable segmention maps. On the Donald Trump and Nicolas Cage data set, our method performs similarly to that of DeepFake. More testing is required to compare the robustness of the methods on other datasets.

[7] https://youtu.be/dV2q3ncXuRM
[8] https://github.com/deepfakes/faceswap
[9] https://github.com/YuvalNirkin/face_swap

5 Conclusion

MoreMori aims to swap the face of Dr. Mori onto any face in real time. To achieve real time face transformations, *MoreMori* uses a cycle consistent GAN trained on specially prepared data. Preparation of the training data offloads the work of a segmenter into the *CycleGAN* model, reducing the computing resources required to evaluate the model for multiple faces in real time. A three stage transfer learning process was utilized to overcome low quality training data available for Dr. Mori.

Experiments suggest that when datasets lacking diversity are used, our method yields qualitative improvement in results over *CycleGAN*. Our method does not use facial specific features, and the face segmenter used to generate masks could be substituted for another segmenter trained for use in the target domain, making the method general.

We were able to achieve our goal of using a GAN based method to produce a real time demonstration of face transfer with multiple faces swapped simultaneously on a consumer laptop. Although difficult to quantify, qualitative feedback from viewers of the demo were able to identify the resemblance of Dr. Mori in the generated images.

References

[1] Nirkin, Y., Masi, I., Tran, A.T., Hassner, T., & Medioni, G. (2018) On face segmentation, face swapping, and face perception. In *International Conference on Automatic Face and Gesture Recognition.*

[2] Zhu, J., Park, T., Isola, P., & Efros, A.A. (2017) Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision (ICCV).*

[3] Chen, Y., Xu, C., Yang, X., & Tao. D. (2018) Attention-GAN for Object Transfiguration in Wild Images. In *European Conference on Computer Vision (ECCV).*

[4] Mejjati, Y.A., Richardt, C., Tompkin, J., Cosker, D., & Kim, K.I. (2018) Unsupervised Attention-guided Image-to-Image Translation. In *Neural Information Processing Systems (NIPS)*

[5] Viola, P., & Jones, M. (2001) Rapid Object Detection using a Boosted Cascade of Simple Features. In *Conference on Computer Vision and Pattern Recognition*

[6] Liu, M., Breuel, T., & Kautz, J. (2017) Unsupervised Image-to-Image Translation Networks. In *Comuting Research Repository (CoRR)*

[7] Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2018) Self-Attention Generative Adversarial Networks. *arXiv preprint arXiv:1805.08318.*

[8] Zhang, G., Kan, M., Shan, S., & Chen, X. (2018) Generative Adversarial Network with Spatial Attention for Face Attribute Editing. In *European Conference on Computer Vision (ECCV).*

[9] Radford, A. & Metz, L. (2016) Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*

[10] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018) A Survey on Deep Transfer Learning. In *International Conference on Artificial Neural Networks (ICANN 2018).*

[11] Fu, H., Gonng, M., Wang, C., Batmanghelich, K., Zhang, K., & Tao, D. (2018) Geometry-Consistent Generative Adversarial Networks for One-Sided Unsupervised Domain Mapping. In *Neural Information Processing Systems (NIPS)*