Introducing Explicit Gaze Constraints to Face Swapping

Ethan Wilson
ethanwilson@ufl.edu
University of Florida
Gainesville, Florida, USA

Frederick Shic
fshic@uw.edu
University of Washington
Seattle, Washington, USA

Eakta Jain
ejain@ufl.edu
University of Florida
Gainesville, Florida, USA

ABSTRACT

Face swapping combines one face's identity with another face's non-appearance attributes (expression, head pose, lighting) to generate a synthetic face. This technology is rapidly improving, but falls flat when reconstructing some attributes, particularly gaze. Image-based loss metrics that consider the full face do not effectively capture the perceptually important, yet spatially small, eye regions. Improving gaze in face swaps can improve naturalness and realism, benefiting applications in entertainment, human computer interaction, and more. Improved gaze will also directly improve Deepfake detection efforts, serving as ideal training data for classifiers that rely on gaze for classification. We propose a novel loss function that leverages gaze prediction to inform the face swap model during training and compare against existing methods. We find all methods to significantly benefit gaze in resulting face swaps.

KEYWORDS

face swapping, gaze prediction, deep learning

ACM Reference Format:
Ethan Wilson, Frederick Shic, and Eakta Jain. 2023. Introducing Explicit Gaze Constraints to Face Swapping. In *2023 Symposium on Eye Tracking Research and Applications (ETRA '23), May 30-June 2, 2023, Tubingen, Germany*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3588015.3588416

1 INTRODUCTION

Face swapping is the act of placing a *character's* face overtop of an *original* face in a piece of media. In deep learning, face swapping is distinct from the face generation task — the network needs to create a realistic face that preserves the original attributes (such as head pose, gaze direction, and mouth movements) while having explicit control over the face's identity. Current face swapping methods have solved the identity reconstruction task, but generally match all other attributes of the face using a black-box approach.

We leverage a pretrained gaze estimation network to optimize an existing face swapping pipeline. Using predicted gaze angles of original and reconstructed faces, we define a reconstruction loss term focused on the eyes to add a gaze component to the overall optimization function, enhancing the accuracy of reconstructed gaze without compromising visual fidelity. Our explicit focus on preserving gaze behavior could be applied to future face swapping pipelines. Our implementation alters the optimization function but does not alter model architecture, meaning that already-trained models can be fine-tuned with this improvement in place. We showcase the method on a popular open-source face swapping network, seeing significant improvement in reconstructed gaze directions compared to baseline face swapping.

1.1 Main Contribution

The proposed design improves upon the naturalness and correctness of gaze behavior in generated face swaps, incorporating a pretrained deep-learning network to guide training in a novel way. Our methodology and experiments provide an implementation guide for facial attribute-based loss functions and reveal their effectiveness, respectively.

1.2 Ethics of Face Swapping

Face swapping has uses in visual effects, interactions with virtual avatars [Caporusso 2021; Foreman 2019], and privacy protection [Lee et al. 2021; Wilson et al. 2022; Zhu et al. 2020]; however, face swapping has become a controversial technology due to its potential for impersonation, spreading misinformation and violating individuals' privacy. These so called *Deepfakes*' sudden accessibility has incited public concern and sparked legislative response [Wagner and Blewer 2019]. Yet, responsible innovation on face swapping is necessary and will lead to positive outcomes. The methods this paper explores can increase the naturalness of future face swapping algorithms, making them more feasible in positive applications for social good. These innovations will also aid in the detection of Deepfakes. Classifiers based on biometric signals, including gaze patterns, are being developed for Deepfake detection [Ciftci et al. 2020a,b; Demir and Ciftci 2021; Jung et al. 2020; Li et al. 2018]. These methods train on real and swapped face videos. By feeding these models new training data with more believable gaze, we will see increased accuracy and reliability when detecting fake media across the internet.

2 RELATED WORK

Recent innovations in image generation techniques, most prominently the generative adversarial network (GAN) [Goodfellow et al. 2014], variational autoencoder (VAE) [Kingma and Welling 2013] and improvements thereafter [Hou et al. 2017; Karras et al. 2019, 2020; Liu and Tuzel 2016; Radford et al. 2016; Razavi et al. 2019], have rapidly advanced the ability to create realistic AI-synthesized faces. These technologies paved the way for powerful, fully automated face swaps that have become nearly undetectable to naive human viewers.

The original image-based face swapping algorithm [deepfakes 2017] is a forked autoencoder with two distinct decoders, each training on a unique identity. Advancements over this initial method have focused on swapping between arbitrary identities [Chen et al.

ETRA '23, May 30-June 2, 2023, Tubingen, Germany

2020; Li et al. 2019; Nirkin et al. 2019], real-time applications [Korshunova et al. 2017], and achieving higher resolutions [Naruniec et al. 2020; Zhu et al. 2021].

Some image and face synthesis methods have begun to leverage existing networks, hereafter referred to as *pretrained expert models*, as part of their training process. Facial recognition systems [Deng et al. 2019] have been used to automatically segment identity or to obtain an overall attribute profile [Chen et al. 2020; Korshunova et al. 2017; Nitzan et al. 2020; Tang et al. 2019]; facial attribute extractors have been used to classify the face in an unsupervised manner [Li and Lin 2019]; landmark estimators have been used to extract or enforce body/facial structure [Kuang et al. 2021; Nitzan et al. 2020; Siarohin et al. 2021; Sun et al. 2018]; style transfer algorithms extract style using pretrained networks' intermediate features [Gatys et al. 2016; Johnson et al. 2016; Liu et al. 2021; Zhang and Dana 2018]. These methods found success using high-level predictions from pretrained expert models to aid in training without requiring supervised labels.

The core goal of modern face swapping is to disentangle the embedded feature vector between identity and other facial attributes, so that identities can be swapped while all other features remain constant. While each algorithm is unique, in nearly all methods the problem is framed as **identity** *versus* **all attributes**, i.e. all aspects outside of identity are placed under a single loss term. For example, multiple approaches isolate and replace the identity portion within an autoencoder's feature embedding [Chen et al. 2020; Korshunova et al. 2017; Li et al. 2019; Wang et al. 2021]. The overall attribute profile is preserved, but is enforced only according to a general image-based reconstruction loss, which may fail to emphasize perceptually relevant features. Particularly, the eyes spatially occupy only about 5.6% of the face, yet human viewers focus on the eyes approximately 40% of the time [Janik et al. 1978]. Because features are derived implicitly from pixel images, the eyes are not prioritized, thus have been found to account for a large percent of noticed artifacts [Wöhler et al. 2021].

A simple way to improve results is the brute-force approach — create a deeper network with a larger latent space. For example, using eight identity-specific decoders rather than two and increasing model depth [Naruniec et al. 2020]. This is effective yet sees increased training times and memory requirements. Instead of increasing resources, another potential solution is to add a gaze-aware constraint to the training process, but this method is not well explained or evaluated in the corresponding manuscript [Perov et al. 2021].

This work details methods to impose explicit constraints on the facial attribute profile, incentivizing the network to better preserve the behavior of the eyes. Our proposed method is modular and could easily extend existing face swapping architectures, leveraging pretrained expert models to better inform models of perceptually important features such as gaze.

3 METHODOLOGY

We propose a novel method to explicitly prioritize gaze over all other implicitly defined facial attributes when training face swapping models. The proposed method leverages a pretrained gaze estimation network, using the resulting gaze values to formulate a reconstruction loss focused on the eye region of the face. Our approach is flexible and generalizes, meaning that it can be applied to any face swapping architecture and with other pretrained expert models. Already-trained models could also be fine-tuned with this improvement. We evaluate our method on DeepFaceLab (DFL) [Perov et al. 2021], comparing against both a gaze-unaware baseline model and their native solution, which had not been formally analyzed or explained in the literature.

3.1 Overview of DeepFaceLab

DFL is the most popular publicly available face swapping platform, so is representative of the majority of face swaps found online. There are many resources online to aid in understanding DFL's pipeline[1]. For explanation and justification of DFL's model design, please refer to their manuscript[2].

We use DFL's LIAE architecture, which disentangles identity with intermediate networks between the encoder and decoder (see Figure 1). The first intermediate network I_{AB} generates latent vectors z^{AB}_{char} and z^{AB}_{orig} during the training process. The second intermediate network I_B is only given the original identity to generate z^B_{orig}. Before passing to the decoder, the latent vectors are concatenated: the original face's becomes $z^{AB}_{orig}||z^B_{orig}$ and the character face's concatenates a copy of itself to become $z^{AB}_{char}||z^{AB}_{char}$. These latent vectors are passed through the respective decoders to reconstruct the input faces. During face swapping, the original face is only passed through I_{AB} and concatenated onto itself to generate latent code $z^{AB}_{orig}||z^{AB}_{orig}$, which is then fed through the decoder to generate a face swapped result.

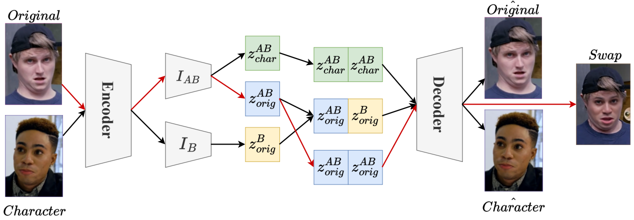


Figure 1: Illustration of DFL's LIAE architecture. The pathway taken to create the resulting face swap is displayed in red. Note that z^{AB}_{char} is concatenated with a copy of itself to reconstruct the character face, and z^{AB}_{orig} is concatenated with itself to produce the face swap result.

Intuitively, we can interpret the first latent vector z_1 to contain attributes, the second vector z_2 to contain identity information, and $z_1||z_2$ to contain full facial information. The latent vector z^{AB} never represents the original face's identity during training, so becomes hardwired to the character face's identity. The LIAE design can be seen in Figure 1.

During training, DFL uses segmentation masks to isolate the error calculation to relevant parts of the face [Bulat and Tzimiropoulos 2017]. The three masks utilized are of the face (M_{face}), the eyes

[1] https://mrdeepfakes.com/forums/thread-guide-deepfacelab-2-0-guide
[2] https://arxiv.org/abs/2005.05535

(M_{eyes}), and the eyes plus mouth (M_{em}). In the following equations, we define the input faces as Y and their reconstructions as $\hat{Y}$[3]. The reconstruction loss combines difference of structural similarity (DSSIM) [Wang et al. 2004; Zhao et al. 2017] and mean squared error (MSE). DSSIM enforces structural consistency between the input and output face using luminance, contrast, and structural components, and MSE error enforces pixel-wise similarity. The loss equations are as follows:

$$SSIM(Y,\hat{Y}) = \frac{(2\mu_i\mu_j + c_1)(2\sigma_{ij} + c_2)}{(\mu_i^2 + \mu_j^2 + c_1)(\sigma_i^2 + \sigma_j^2 + c_2)} \tag{1}$$

$$L_{DSSIM}(Y,\hat{Y}) = \frac{1 - SSIM(Y,\hat{Y})}{2} \tag{2}$$

where i, j = sliding windows of size NxN
μ_i, μ_j = average of i, j $\quad\sigma_i^2, \sigma_j^2$ = variance of i, j
σ_{ij} = covariance of i, j $\quad c_1, c_2$ = stabilizing variables

$$L_{MSE}(Y,\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \tag{3}$$

The core reconstruction loss is a weighted sum between DSSIM, MSE, and an MSE calculation comparing the input and predicted face masks:

$$L_{\triangle}(Y,\hat{Y},M_{face},\hat{M_{face}}) = \lambda_1 L_{DSSIM}(Y,\hat{Y}) + \lambda_2 L_{MSE}(Y,\hat{Y}) + \lambda_3 L_{MSE}(M_{face},\hat{M_{face}}) \tag{4}$$

DFL can explicitly target facial attributes via its optional eyes and mouth priority term. This integrates well with the main loss equation, measuring the absolute value of pixel error between the original and generated faces masked to the eyes and mouth. This is an optional term that must be enabled by DFL users.

$$L_{\triangle em}(Y,\hat{Y},M_{em}) = \lambda_{em}|YM_{em} - \hat{Y}M_{em}| \tag{5}$$

3.2 Proposed Gaze Reconstruction Loss

Motivated by previous image generation methods' success using pretrained expert models, we leverage a gaze estimation network. We incorporate L2CS-Net[4] [Abdelrahman et al. 2022], which predicts pitch and yaw angles μ, ϕ from input face images. This network is optimized towards unconstrained environments so is well suited to the data typical in training face swaps. We incentivize the face swapping model to better reconstruct gaze by penalizing offsets in predicted gaze angle between the input and reconstructed faces during training.

Our gaze reconstruction loss is computed as follows. μ and ϕ are converted to normalized Cartesian coordinates, then the angle θ between the two vectors is found.

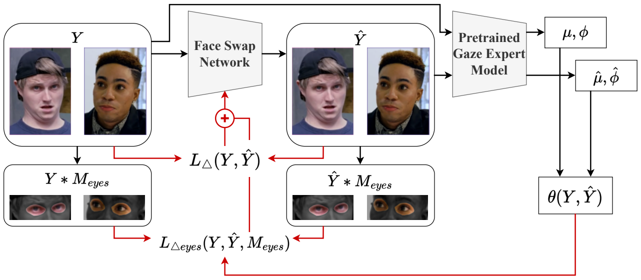


Figure 2: Design diagram of the steps to compute the gaze reconstruction loss.

$$\begin{aligned} \mu,\phi &= L2CS(Y) & \hat{\mu},\hat{\phi} &= L2CS(\hat{Y}) \\ x &= sin(\phi)cos(\mu) & y &= sin(\phi)sin(\mu) \\ z &= cos(\phi) & V_1 &=< x,y,z > \\ \hat{x} &= sin(\hat{\phi})cos(\hat{\mu}) & \hat{y} &= sin(\hat{\phi})sin(\hat{\mu}) \\ \hat{z} &= cos(\hat{\phi}) & V_2 &=< \hat{x},\hat{y},\hat{z} > \end{aligned}$$

$$\text{Error is computed as:} \quad \theta(Y,\hat{Y}) = cos^{-1}\left(\frac{V_1 \cdot V_2}{||V_1||||V_2||}\right) \tag{6}$$

We apply this error term only to the regions of the network that correspond to the eyes. We use Y, $\hat{Y}$, and M_{eyes} to construct a reconstruction loss specific to the eyes that can be scaled by the computed θ and hyperparameters α and β. We structure our loss equation similarly to equation 4, using DSSIM and MSE computations on the original and reconstructed image eye regions. An illustration of the design and steps taken to compute the loss equation can be seen in Figure 2.

$$L_{\triangle eyes}(Y,\hat{Y},M_{eyes}) = \theta(Y,\hat{Y})\Big(\alpha L_{DSSIM}(YM_{eyes},\hat{Y}M_{eyes}) + \beta L_{MSE}(YM_{eyes},\hat{Y}M_{eyes})\Big) \tag{7}$$

4 EVALUATION

We assess the performance of each condition by analyzing the offset in viewing angles between the face swap and the real face in the corresponding source video. To compute this metric, we utilize L2CS-Net [Abdelrahman et al. 2022] to predict a gaze viewing angle for each condition, considering the source video's predicted gaze vector to be the ground truth. In our evaluation we use DFL's set parameters for our λ values. Namely, $\lambda_1, \lambda_2, \lambda_3 = 10$, $\lambda_e m = 300$. When implementing our proposed loss term, we use $\alpha = 3$ and $\beta = 30$.

4.1 Dataset

We introduce a dataset to serve as a testing ground for our approach. We generate our face swaps using the source video clips taken from the FaceForensics++ Deep Fake Detection Dataset[5] [Rossler et al.

[3] Note that original and character faces' reconstruction loss are computed in identical fashion.
[4] https://github.com/Ahmednull/L2CS-Net
[5] https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

2019]. In the dataset, subjects perform the same tasks[6], ensuring similar expression and head pose, making these clips ideal for high quality face swaps. Our dataset consists of six subject (three female, three male). For each gender, two subjects have similar appearance to one another. Per gender, we permute all combinations of subjects being used as the character and original face, resulting in a total of 12 unique face pairs.

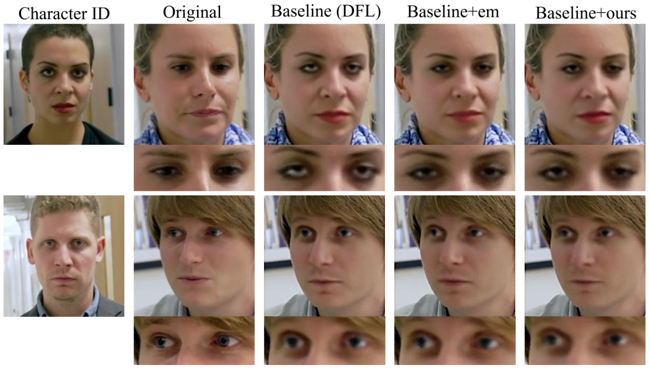


Figure 3: Visual comparison of face swaps produced by the baseline DFL method, DFL with eyes and mouth priority loss (em), and DFL with our proposed gaze loss. Both improvements over the baseline reduce gaze angle error.

We generate face swaps across multiple conditions, keeping all other hyperparameters consistent. Every model is pretrained for 100 thousand iterations on the CelebA dataset [Liu et al. 2015], then trained for the final 20 thousand iterations on the identity pair. Frames from our generated dataset can be seen in Figure 3. The conditions are:

- **DFL.** The model implicitly learns gaze behavior while optimizing the core reconstruction loss in equation 4.
- **DFL+em.** DeepFaceLab with eyes and mouth priority loss enabled. DFL's native solution which further enforces pixel-wise similarity for the key regions of the face.
- **DFL+Gaze.** DeepFaceLab with our proposed gaze loss. The model explicitly enforces consistency using gaze vectors computed by the pretrained expert model.
- **DFL+Gaze (finetuning).** The model is pretrained with no gaze-specific loss, then trained for the final 20 thousand iterations using our proposed loss.
- **DFL+em+Gaze.** Both DFL's native approach and our proposed approach are enabled during training.

4.2 Results

We analyze error values, collapsing from individual frames (∼ 2900 per video) to average across each individual in the dataset. The baseline DFL produces an average error of 5.98°[95% Confidence Interval (CI): 4.82, 7.13]. All improvements on the baseline method produce noticeably more accurate gaze values: DFL+em averages 4.85°[95% CI: 3.80, 5.90], DFL+Gaze averages 4.71°[95% CI: 3.66, 5.77], DFL+Gaze (finetuning) averages 4.85°[95% CI: 3.80, 5.90], and DFL+em+Gaze averages 4.72°[95% CI: 3.67, 5.77]. On the test dataset, introducing DFL's eyes and mouth priority term decreases reconstructed gaze error by 18.1%; introducing the proposed method decreases by 19.7%, and introducing both components decreases gaze error by 20.32%.

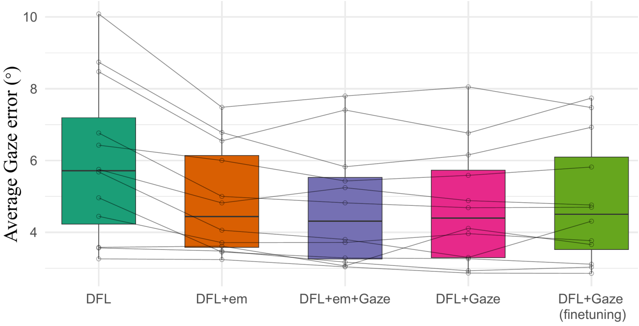


Figure 4: Plot of mean gaze error across all evaluated videos ($N = 12$) by condition. Individual video results are plotted over-top and connected across each box plot.

We test for significance via a linear mixed-effects model. We first compute the average of the log of angular error for each method and individual, applying the log transform to improve normality of error distributions. We then model errors as average(log(error)) method with a random intercept per individual. All improved methods significantly improve over DFL ($p < 0.001$). However, we have not found statistical evidence pair-wise between any of the improved methods. Interestingly, the DFL+em+Gaze approach combining pixel information and explicit gaze modeling yielded insignificant benefit over DFL+em ($\text{t}(1, 44) = 1.603, p = 0.116$). This may indicate that the two optimizations capture similar underlying information.

Each method's performance across individuals in the dataset is plotted in Figure 4. We see a large amount of variability among individual video results in all methods other than DFL, indicating roughly equivalent performance for all improvements analyzed. Looking on an individual video basis (Figure 5), relative error remains quite stable across the dataset, suggesting that error is tied to the properties of the video, i.e. the specific pair of faces involved.

5 DISCUSSION & CONCLUSION

Based on our experiments, the proposed gaze improvement for face swapping using a pretrained gaze prediction model largely decreases gaze error by 19.7% when appended to an image-based reconstruction loss equation. This analysis provided key information in enhancing gaze behavior in face swapping models and the potential benefit that can be provided.

Our method achieves similar performance compared to DFL's native solution to the problem (which had not previously been quantified relative to the baseline). These adjunct approaches likely capture the same information. However, it is important to note that the vast majority of face swapping approaches implement **neither** approach, so either will improve gaze representation. Compared to the baseline, a few degrees may or may not have a noticeable

[6] The video segments we use are: exit phone room, kitchen pan, outside talking pan laughing, walking outside cafe disgusted. These are concatenated into a single video $\tilde{2}$ minutes in length per subject.

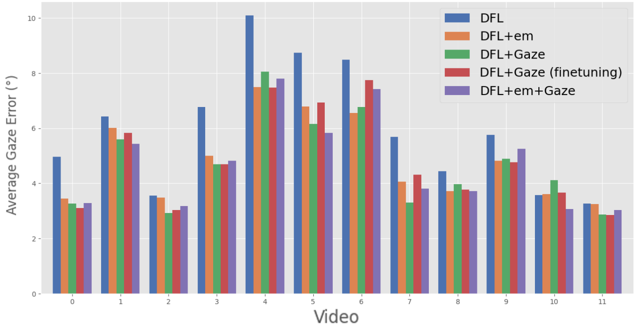


Figure 5: Average gaze error in degrees for all conditions evaluated, plotted by individual video.

impact on viewer perception. However, these improvements could be quite beneficial to aid in the development of biometric Deepfake classifiers that leverage gaze to label video as real or fake.

Our method uses the same pretrained network in training as our evaluation pipeline. This opens up the possibility that our model could have learned to minimize the prediction error for L2CS-Net rather than generally improving gaze representations. However, by observing how visually similar our results are to the native solution and the minimal differences in gaze errors, this concern is alleviated. Our pipeline leveraged the pretrained gaze model to derive an angle error θ. If we had instead granted white-box access to the pretrained model, fitting to the gaze model would be more likely.

Unlike the native approach, our proposed method incorporates gaze angle as a high-level feature. This lessens the dependence on pixel-level matching of the eyes, possibly being more impactful at higher resolutions. While this analysis focused fully on gaze, a similar pipeline could be easily developed for other features, such as expression or head-pose matching. Stacking multiple optimizations on the same network could improve overall fidelity. Analyzing the interaction between multiple pretrained expert models as they guide the same model's training process is a worthwhile future direction. The dependence on pretrained expert models to compute gaze vectors will make our approach more appealing as more advanced predictors are developed. For example, current gaze predictors are prone to around 4 degrees of prediction error, which is likely acting as a lower bound on our method's performance. When better performing predictors are created, our system will improve accordingly.

In this paper, we presented a novel loss component that significantly increases a face swapping model's ability to accurately reconstruct gaze. We compared multiple design decisions, including a formal analysis of DFL's eyes and mouth priority method. Our most successful implementation, combining both optimizations, decreased gaze reconstruction error by 20.32%. This advancement improves face swapping technology but is particularly promising for gaze-based Deepfake detection; such an increase in fidelity will allow researchers to generate higher quality training datasets that will lead to better Deepfake detection in real-world settings.

REFERENCES

Ahmed A. Abdelrahman, Thorsten Hempel, Aly Khalifa, and Ayoub Al-Hamadi. 2022. L2CS-Net: Fine-Grained Gaze Estimation in Unconstrained Environments. https://doi.org/10.48550/ARXIV.2203.03339

Adrian Bulat and Georgios Tzimiropoulos. 2017. How Far Are We From Solving the 2D & 3D Face Alignment Problem? (And a Dataset of 230,000 3D Facial Landmarks). In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Nicholas Caporusso. 2021. Deepfakes for the Good: A Beneficial Application of Contentious Artificial Intelligence Technology. In *Advances in Artificial Intelligence, Software and Systems Engineering (Advances in Intelligent Systems and Computing)*, Tareq Ahram (Ed.). Springer International Publishing, Cham, 235–241. https://doi.org/10.1007/978-3-030-51328-3_33

Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. 2020. SimSwap: An Efficient Framework For High Fidelity Face Swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 2003–2011. https://doi.org/10.1145/3394171.3413630

Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. 2020a. FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 1–1. https://doi.org/10.1109/TPAMI.2020.3009287

Umur Aybars Ciftci, İlke Demir, and Lijun Yin. 2020b. How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*. 1–10. https://doi.org/10.1109/IJCB48548.2020.9304909

deepfakes. 2017. deepfakes_faceswap. https://github.com/deepfakes/faceswap

Ilke Demir and Umur Aybars Ciftci. 2021. Where Do Deep Fakes Look? Synthetic Face Detection via Gaze Tracking. In *ACM Symposium on Eye Tracking Research and Applications* (Virtual Event, Germany) *(ETRA '21 Full Papers)*. Association for Computing Machinery, New York, NY, USA, Article 6, 11 pages. https://doi.org/10.1145/3448017.3457387

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alison Foreman. 2019. Salvador Dalí deepfake brings legendary surrealist to life at Florida museum. https://mashable.com/article/salvador-dali-deepfake Section: Tech.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image Style Transfer Using Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu. 2017. Deep Feature Consistent Variational Autoencoder. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1133–1141. https://doi.org/10.1109/WACV.2017.131

Stephen W. Janik, A. Rodney Wellens, Myron L. Goldberg, and Louis F. Dell'Osso. 1978. Eyes as the Center of Focus in the Visual Examination of Human Faces. *Perceptual and Motor Skills* 47, 3 (1978), 857–858. https://doi.org/10.2466/pms.1978.47.3.857 arXiv:https://doi.org/10.2466/pms.1978.47.3.857 PMID: 740480.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*. Springer, 694–711.

Tackhyun Jung, Sangwon Kim, and Keecheon Kim. 2020. DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. *IEEE Access* 8 (2020), 83144–83154. https://doi.org/10.1109/ACCESS.2020.2988660

Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4396–4405. https://doi.org/10.1109/CVPR.2019.00453 ISSN: 2575-7075.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8107–8116. https://doi.org/10.1109/CVPR42600.2020.00813 ISSN: 2575-7075.

Diederik P Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. https://doi.org/10.48550/ARXIV.1312.6114

Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. 2017. Fast Face-Swap Using Convolutional Neural Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Zhenzhong Kuang, Huigui Liu, Jun Yu, Aikui Tian, Lei Wang, Jianping Fan, and Noboru Babaguchi. 2021. Effective De-identification Generative Adversarial Network for Face Anonymization. In *Proceedings of the 29th ACM International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 3182–3191. https://doi.org/10.1145/3474085.3475464

Sooyeon Lee, Abraham Glasser, Becca Dingman, Zhaoyang Xia, Dimitris Metaxas, Carol Neidle, and Matt Huenerfauth. 2021. American Sign Language Video Anonymization to Support Online Participation of Deaf and Hard of Hearing Users. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. Number 22. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3441852.3471200

Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2019. FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping. https://doi.org/10.48550/ARXIV.1912.13457

Tao Li and Lei Lin. 2019. AnonymousNet: Natural Face De-Identification With Measurable Privacy. 56–65. https://doi.org/10.1109/CVPRW.2019.00013 ISSN: 2160-7516.

Yuezun Li, Ming-Ching Chang, and Siwei Lyu. 2018. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. 1–7. https://doi.org/10.1109/WIFS.2018.8630787

Mingcong Liu, Qiang Li, Zekui Qin, Guoxin Zhang, Pengfei Wan, and Wen Zheng. 2021. BlendGAN: Implicitly GAN Blending for Arbitrary Stylized Face Generation. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 29710–29722. https://proceedings.neurips.cc/paper/2021/file/f8417d04a0a2d5e1fb5c5253a365643c-Paper.pdf

Ming-Yu Liu and Oncel Tuzel. 2016. Coupled Generative Adversarial Networks. In *Advances in Neural Information Processing Systems*, Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/hash/502e4a16930e414107ee22b6198c578f-Abstract.html

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

J. Naruniec, L. Helminger, C. Schroers, and R.M. Weber. 2020. High-Resolution Neural Face Swapping for Visual Effects. *Computer Graphics Forum* 39, 4 (2020), 173–184. https://doi.org/10.1111/cgf.14062 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14062

Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019. FSGAN: Subject Agnostic Face Swapping and Reenactment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. 2020. Face Identity Disentanglement via Latent Space Mapping. *ACM Trans. Graph.* 39, 6, Article 225 (Nov 2020), 14 pages. https://doi.org/10.1145/3414685.3417826

Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. 2021. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *arXiv:2005.05535 [cs, eess]* (June 2021). http://arxiv.org/abs/2005.05535 arXiv: 2005.05535.

Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv:1511.06434 [cs]* (Jan. 2016). http://arxiv.org/abs/1511.06434 arXiv: 1511.06434.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. 2019. Generating Diverse High-Fidelity Images with VQ-VAE-2. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/hash/5f8e2fa1718d1bbcadf1cd9c7a54fb8c-Abstract.html

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Aliaksandr Siarohin, Oliver J. Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. 2021. Motion Representations for Articulated Animation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13648–13657. https://doi.org/10.1109/CVPR46437.2021.01344 ISSN: 2575-7075.

Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. 2018. Natural and Effective Obfuscation by Head Inpainting. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5050–5059. https://doi.org/10.1109/CVPR.2018.00530 ISSN: 2575-7075.

Hao Tang, Dan Xu, Gaowen Liu, Wei Wang, Nicu Sebe, and Yan Yan. 2019. Cycle In Cycle Generative Adversarial Networks for Keypoint-Guided Image Generation. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*. Association for Computing Machinery, New York, NY, USA, 2052–2060. https://doi.org/10.1145/3343031.3350980

Travis L. Wagner and Ashley Blewer. 2019. "The Word Real Is No Longer Real": Deepfakes, Gender, and the Challenges of AI-Altered Video. *Open Information Science* 3, 1 (2019), 32–46. https://doi.org/doi:10.1515/opis-2019-0003

Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. 2021. HifiFace: 3D Shape and Semantic Prior Guided High Fidelity Face Swapping. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 1136–1142. https://doi.org/10.24963/ijcai.2021/157 Main Track.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. https://doi.org/10.1109/TIP.2003.819861

Ethan Wilson, Frederick Shic, Jenny Skytta, and Eakta Jain. 2022. Practical Digital Disguises: Leveraging Face Swaps to Protect Patient Privacy. *arXiv:2204.03559 [cs]* (April 2022). http://arxiv.org/abs/2204.03559 arXiv: 2204.03559.

Leslie Wöhler, Martin Zembaty, Susana Castillo, and Marcus Magnor. 2021. Towards Understanding Perceptual Differences between Genuine and Face-Swapped Videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 240, 13 pages. https://doi.org/10.1145/3411764.3445627

Hang Zhang and Kristin Dana. 2018. Multi-style Generative Network for Real-time Transfer. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.

Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. 2017. Loss Functions for Image Restoration With Neural Networks. *IEEE Transactions on Computational Imaging* 3, 1 (2017), 47–57. https://doi.org/10.1109/TCI.2016.2644865

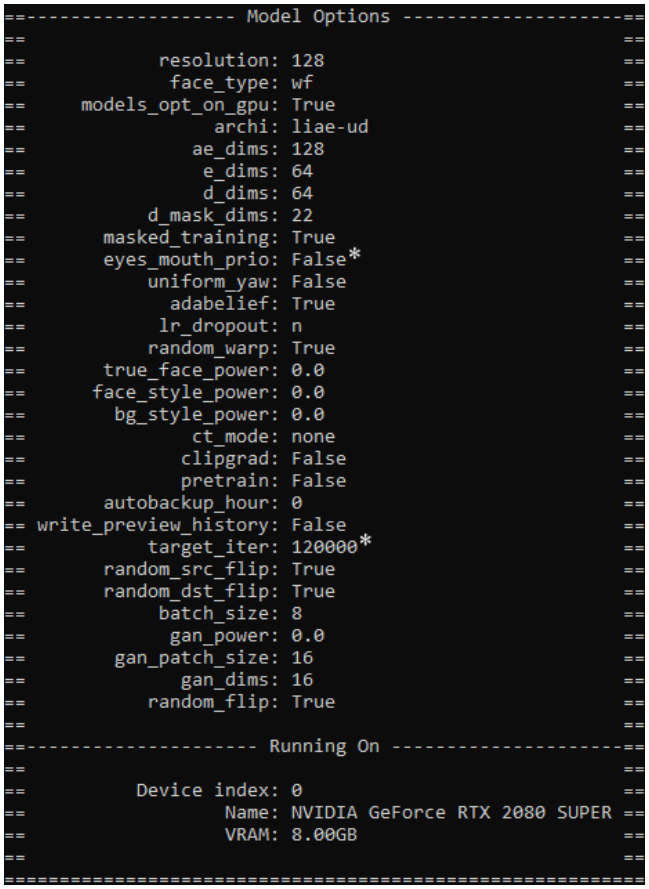
Bingquan Zhu, Hao Fang, Yanan Sui, and Luming Li. 2020. Deepfakes for Medical Video De-Identification: Privacy Protection and Diagnostic Information Preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, USA, 414–420. https://doi.org/10.1145/3375627.3375849

Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. 2021. One Shot Face Swapping on Megapixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4834–4844.

A SUPPLEMENTARY INFORMATION

A.1 DFL Parameters

All face swaps were generated using a NVIDIA GeForce RTX 2080 Super. The exact run configurations used to generate our face swaps are outlaid. For training parameters, refer to Figure 6. For merge parameters, refer to Table 1.



```
==------------------- Model Options --------------------==
==                                                        ==
==            resolution: 128                             ==
==             face_type: wf                              ==
==     models_opt_on_gpu: True                            ==
==                 archi: liae-ud                         ==
==               ae_dims: 128                             ==
==                e_dims: 64                              ==
==                d_dims: 64                              ==
==           d_mask_dims: 22                              ==
==       masked_training: True                            ==
==       eyes_mouth_prio: False*                          ==
==           uniform_yaw: False                           ==
==             adabelief: True                            ==
==            lr_dropout: n                               ==
==           random_warp: True                            ==
==       true_face_power: 0.0                             ==
==      face_style_power: 0.0                             ==
==        bg_style_power: 0.0                             ==
==               ct_mode: none                            ==
==              clipgrad: False                           ==
==              pretrain: False                           ==
==       autobackup_hour: 0                               ==
== write_preview_history: False                           ==
==           target_iter: 120000*                         ==
==       random_src_flip: True                            ==
==       random_dst_flip: True                            ==
==            batch_size: 8                               ==
==             gan_power: 0.0                             ==
==        gan_patch_size: 16                              ==
==              gan_dims: 16                              ==
==           random_flip: True                            ==
==                                                        ==
==-------------------- Running On ----------------------==
==                                                        ==
==          Device index: 0                               ==
==                  Name: NVIDIA GeForce RTX 2080 SUPER  ==
==                  VRAM: 8.00GB                          ==
==                                                        ==
============================================================
```

Figure 6: Training parameters used to when generating all face swap stimuli. Note that the LIAE architecture is classified as a SAEHD model in DFL's configurations. The eyes_mouth_prio parameter was set to true when the proposed gaze term was disabled. target_iter varied depending on training phase (100k iterations pretraining, 20k iterations on the end pair of identities).

A.2 Analyzing Distribution of Gaze

We hypothesized that although the DFL+em and DFL+Gaze conditions had similar errors compared to the baseline, the differing approaches may distribute the data in differing manners. By plotting the pitch and yaw vectors across the frames of video segments (Figure 7), we see some differences across conditions. It is visually clear that the baseline DFL distribution does not match the source, and appears to be much more aligned to the horizontal and vertical axes than all other conditions. The other conditions are much closer in appearance, yet appear to have variations both in outlier distribution and cluster shapes. For example, the main cluster for DFL+em in the bottom row appears to match the source best along the yaw axis. However, DFL+Gaze better matches pitch and best mimics the left peninsula that juts out from the bottom of the cluster.

Parameter	Value
mode	(1) overlay
mask mode	(4) learned-prd*learned-dst
erode mask modifier	20
blur mask modifier	80
motion blur power	0
output face scale modifier	0
color transfer to predicted face	rct
sharpen mode	(0) None
super resolution power	0
image degrade by denoise power	0
image degrade by bicubic rescale power	0
degrade color power of final image	0
number of workers	12

Table 1: Merge parameters used to generate our face swap stimuli after training. Note that the original video clips were 1920x1080 pixels at 24 frames per second and faces were extracted at 512x512 pixels.

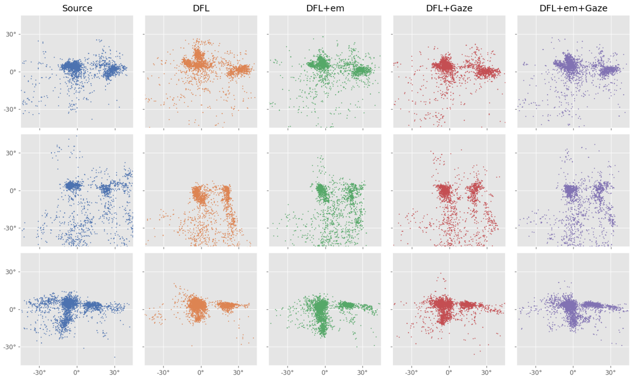


Figure 7: Gaze vectors plotted over all frames of three videos from our dataset (each row corresponding to one full video). Pitch angles plotted on the horizontal axis and yaw on the vertical axis.