Supplementary Material
Contrastive Learning on Synthetic Videos for GAN Latent Disentangling

Kevin Duarte
Adobe Inc. (ASML)
`kduarte@adobe.com`

Wei-An Lin
Adobe Inc. (ASML)
`wlin@adobe.com`

Ratheesh Kalarot
Adobe Inc. (ASML)
`kalarot@adobe.com`

Jingwan Lu
Adobe Research
`jlu@adobe.com`

Eli Shechtman
Adobe Research
`elishe@adobe.com`

Shabnam Ghadar
Adobe Inc. (ASML)
`ghadar@adobe.com`

Mubarak Shah
Center for Research in Computer Vision
University of Central Florida
`shah@crcv.ucf.edu`

In this supplementary material, we include additional diagrams which describe the contrastive learning procedure. Also, we give an in-depth description on how our approach is applied to expression transfer and motion transfer. We also present additional ablations, qualitative results for our proposed approach, and include a discussion about limitations and potential societal impacts of our method.

1 Contrastive Learning

We propose two contrastive losses, which are depicted in Figure 1. The appearance contrastive loss pulls the representation for faces within the same video together, while pushing representations of different identities apart. The structural contrastive loss pulls the representations from augmented versions of the same frame together, while pushing representations of different frames apart. Although the appearance contrastive loss diagram contains two video sequences, it is computed batch-wise.

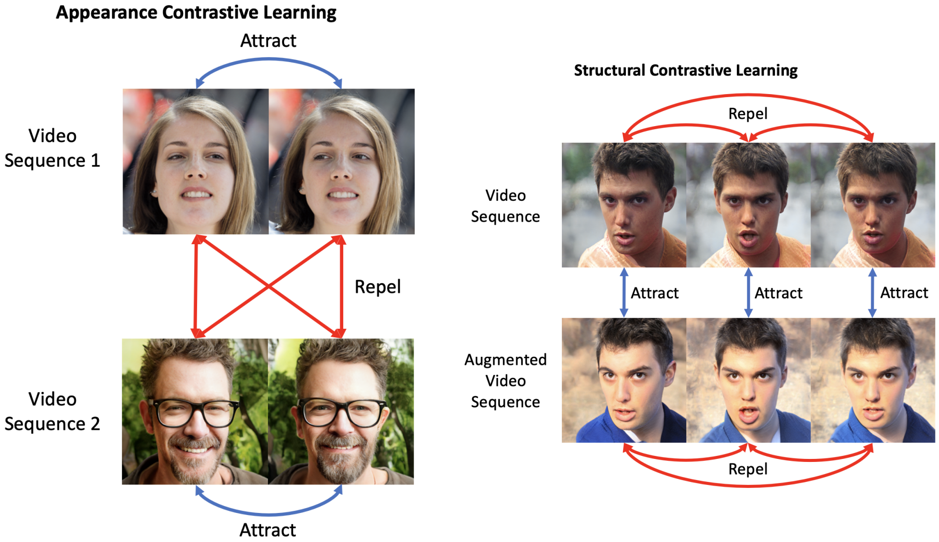

Figure 1: The appearance contrastive loss (left) and structural contrastive loss (right).

NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research.

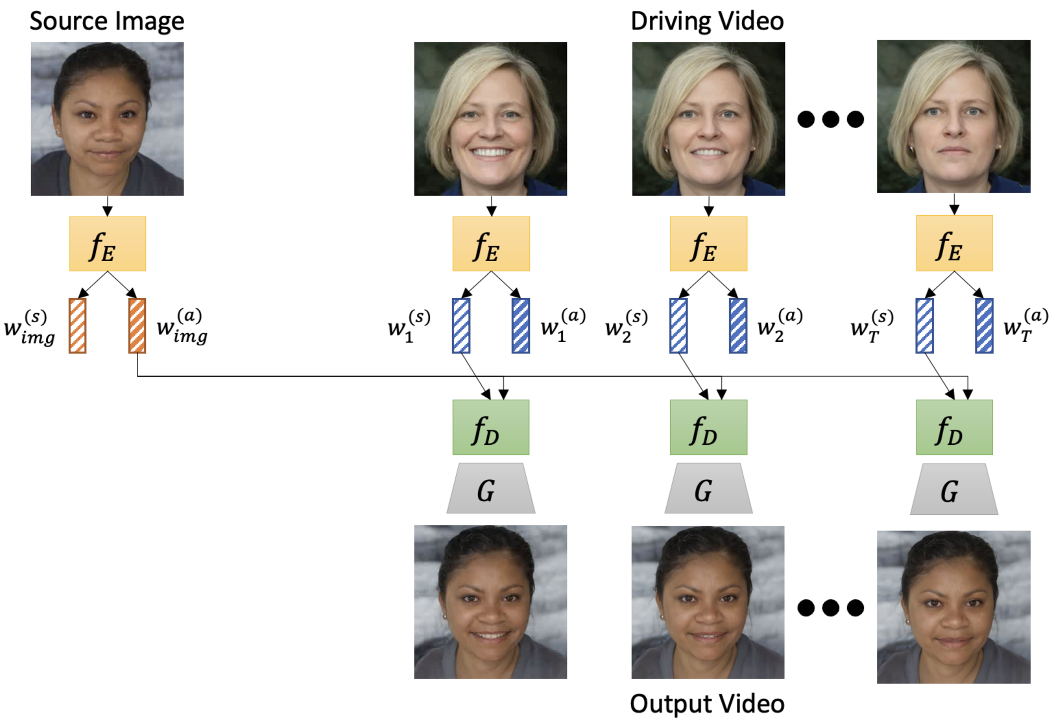

Figure 2: Motion Transfer with our proposed architecture. We extract the appearance and structural features from the source image as well as each frame of the driving video. Then the appearance features of the source image and the structural features of each video frame are used to generate the output video frames.

2 Applications

In this section, we include the technical details about how our method can be applied at inference time for expression and motion transfer.

Expression Transfer As our encoder can disentangle the appearance and structural features of a given latent code, we perform expression transfer by swapping the appearance and structural features of two images and passing them through the decoder. Formally, given two face images x_1 and x_2, we extract their latent representations w_1 and w_2, respectively. If the goal is to generate a face with the identity of x_1 and expression of x_2, we pass both latent codes into our encoder and obtain the appearance features of the first, $w_1^{(a)}$, and the structural features of the second, $w_2^{(s)}$. Then, we generate a new latent code by inputting these representations into the decoder, $\tilde{w} = f_D\left(w_1^{(a)}, w_2^{(s)}\right)$. This code is used by the GAN to generate a realistic face image with the appearance of the first input image and expression of the second.

Motion Transfer We perform motion transfer in a similar fashion, by maintaining the structural features of the driving video and substituting the appearance features of the source image. Given the source image x_{img}, we can extract disentangled features $w_{img}^{(a)}$ and $w_{img}^{(s)}$; likewise, each video frame of the driving video, $\{x_t\}_{t=1}^T$, can be disentangled into $\{w_t^{(a)}\}_{t=1}^T$ and $\{w_t^{(s)}\}_{t=1}^T$. With the decoder, we can obtain the sequence of latent codes $\{f_D(w_{img}^{(a)}, w_t^{(s)})\}_{t=1}^T$, which are used to generate the output video frames. Figure 2 depicts how our method performs motion transfer.

Table 1: Ablations on the VoxCeleb dataset. We evaluate the effect of the auxiliary losses.

Method	No $\mathcal{L}_{3DMM_a}$	No $\mathcal{L}_{3DMM_s}$	No $\mathcal{L}_{cyc-lat}$	No $\mathcal{L}_{cyc-img}$	No L_a nor L_s	Full
AKD	1.517	2.299	1.515	1.976	2.193	1.928
ACED	0.768	0.665	0.730	0.714	0.769	0.711

3 Additional Implementation Details

Hyper-parameter Selection The network is trained for 50 epochs using the Adam optimizer [1] with an initial learning rate of 5e-5 and cosine learning rate scheduler [2]. The margin used in the contrastive losses is $\gamma = 1$, and the weights for each loss are selected empirically: $\lambda_{rec} = 200$, $\lambda_a = 2$, $\lambda_s = 2$, $\lambda_{cyc-lat} = 100$, $\lambda_{cyc-img} = 10$, $\lambda_{3DMM_a} = 0.2$, and $\lambda_{3DMM_s} = 1$. For each batch, we sample 8 frames per video. The batch size for all latent computations is 256, and all losses requiring image generation ($\mathcal{L}_{cyc-img}$, $\mathcal{L}_{3DMM_a}$, and $\mathcal{L}_{3DMM_s}$) have a batch size of 8. Our model is trained on 8 NVIDIA Tesla V100 GPUs.

Evaluation Protocol Quantitatively evaluating the quality of motion transfer is non-trivial since the ground-truth videos are not available. Given a driving video and a source image, we perform motion transfer to obtain an output video. To evaluate how well the motion is transferred, we obtain the Average Keypoint Distance (AKD) between the output and driving videos. These keypoints are extracted using a facial landmark detector pretrained by [3]. As facial landmarks encode identity information (*e.g.* mouth/eye size and face shape can lead to different landmarks for two people with the same expression), we use the same identity for the source image and the driving video; this ensures the metric correctly measures changes in expression and pose, and not identity features. We also evaluate how well the source image's identity is maintained throughout the video by measuring the average classifier embedding distance (ACED) between frames of the output video and the source image. This distance is the $\mathcal{L}_1$ distance between the embedding layer of a ResNet-50 [4] model trained on the UMDFaces dataset [5]. We note, the ACED and AKD metrics should be viewed jointly: whereas ACED measures how well a method maintains a person's identity, AKD measures how well pose and expression are preserved. For this quantitative evaluation, we select 50 driving video sequences[1] from the VoxCeleb2 dataset, each ranging from 100 to 300 frames (4 to 12 seconds). We transfer the motion to faces from selected frames within other videos with the same identities. We also report the Frèchet Inception Distance (FID) [6] and Frèchet Video Distance (FVD) [7] to evaluate each methods' image naturalness and motion quality, respectively.

4 Ablation Experiments

Auxiliary Losses We quantitatively evaluate how the proposed auxiliary losses effect the performance of our method in Table 2. Practically, we aim to find a balance between identity/appearance preservation and producing a face with the correct pose and expression. The 3DMM-based consistency losses lead to predictable changes in the outputs. The network trained without the $\mathcal{L}_{3DMM_a}$ produces suffers from some appearance loss even though it improves in terms of landmark distance; conversely, the network trained without $\mathcal{L}_{3DMM_s}$ maintains the source identity, but does not correctly transfer expressions. We find that the latent cyclic consistency loss ($\mathcal{L}_{cyc-lat}$) leads to improved disentangling: without this loss, the network generates faces with correct expression and pose, but incorrect identities. Lastly, the inclusion of the image-based cyclic consistency loss ($\mathcal{L}_{cyc-img}$) leads to slight improvements in terms of identity preservation and expression transfer. In our final ablation, we show that a network trained without our proposed contrastive losses (No L_a nor L_s) performs poorly on both metrics.

Distance Metric For our proposed method, we use the negative L2 distance to compute the similarity in our contrastive losses (equations 2 and 3). Here, we evaluate the dot product as another similarity metric. We observe that the network trained using the dot product is able to better transfer face pose than the network trained using negative L2 distance. However, the faces produced by this network differ from the target appearance in the source image, with more noticeable changes in

---

[1] All identities and videos used in evaluation are distinct from those used to train our network.

Table 2: Ablations on the VoxCeleb dataset. We evaluate the use of another similarity metric in the contrastive objective (dot product).

Method	AKD↓	ACED↓
Dot product	1.583	0.755
Full Method	1.928	0.711

hairstyle and face shape. This behaviour is reflected in Table 2, where using the dot product leads to improved average keypoint distance, but worse average classifier embedding distance.

5 Discussion

Hyperparameter Selection and Robustness We selected the loss weight hyper-parameters (λ_{rec}, λ_a, λ_s, $\lambda_{cyc-lat}$, $\lambda_{cyc-img}$, λ_{3DMM_a}, and λ_{3DMM_s}) empirically. In selecting the loss weight hyper-parameters, we had two main observations. First, increasing the magnitude of $\mathcal{L}_{rec}$ and $\mathcal{L}_{cyc-lat}$ relative to the other losses led to improved encoder and decoder training. Second, increasing the weight for the 3DMM losses ($\mathcal{L}_{3DMM_a} > 1$ and $\mathcal{L}_{3DMM_s} > 1$) led to a large degradation in performance. In general, the method is robust to small changes in the other loss weights ($\lambda_a = 2$, $\lambda_s = 2$, and $\lambda_{cyc-img} = 10$).

Limitations Although our method can perform expression and motion transfer quite well, we observe that when the appearance of two subjects are vastly different, swapping structural features can sometimes lead to variations in hair style and glasses. Moreover, we find that our approach occasionally struggles with some smaller expression changes, like eye blinking. We attribute this behaviour to the imbalance of training samples where the person has their eyes open vs. closed: since the majority of video frames tend to have the persons' eyes open, there is a lower chance to sample a frame where the eyes are closed.

Societal Impact Our method can be used for a variety of applications. Artists and designers can use our approach to change the expression of faces within images, or to generate realistic videos of a given identity. However, there is always the possibility of our method being nefariously used to generate images or videos for the purpose of spreading “fake news" or propaganda. To prevent this, additional resources should be invested in technologies that can detect fake images [8, 9] and videos [10, 11].

6 Additional Qualitative Results

Attached are 3 motion transfer videos using the Offset Trick, FOMM, and our proposed approach. The driving videos are obtained from the VoxCeleb dataset and the source images are from FFHQ dataset. We include a variety of source face identities and show that our method tends to transfers motion with minimal change in appearance (see Video1.avi and Video2.avi). Moreover, we include additional expression transfer examples in Figure 3.

Failure Cases As mentioned in the discussion, our method can fail when combining the appearance and structural features of different faces. This can result in larger changes in identity as shown in row 2 in Video3.avi. Although our generated videos tend to maintain facial appearance consistent throughout the video, there are instances when certain attributes (*e.g.* hair or glasses) change from frame to frame. This is evident in row 1 of Video3.avi, where the glasses disappear and reappear multiple times within the video. Generally, these failures occur the both the identity and face pose of the source image and driving video are very different.

Qualitative Evaluation of Auxiliary Losses In Figure 4, we show example outputs of our method trained without various auxiliary losses. It can be seen that the 3DMM-based auxiliary losses are complementary - $\mathcal{L}_{3DMM_a}$ and $\mathcal{L}_{3DMM_s}$ aid in maintaining the appearance and structure of the faces respectively. Both cyclic losses, $\mathcal{L}_{cyc-lat}$ and $\mathcal{L}_{cyc-img}$ improve the disentangling ability of the model.

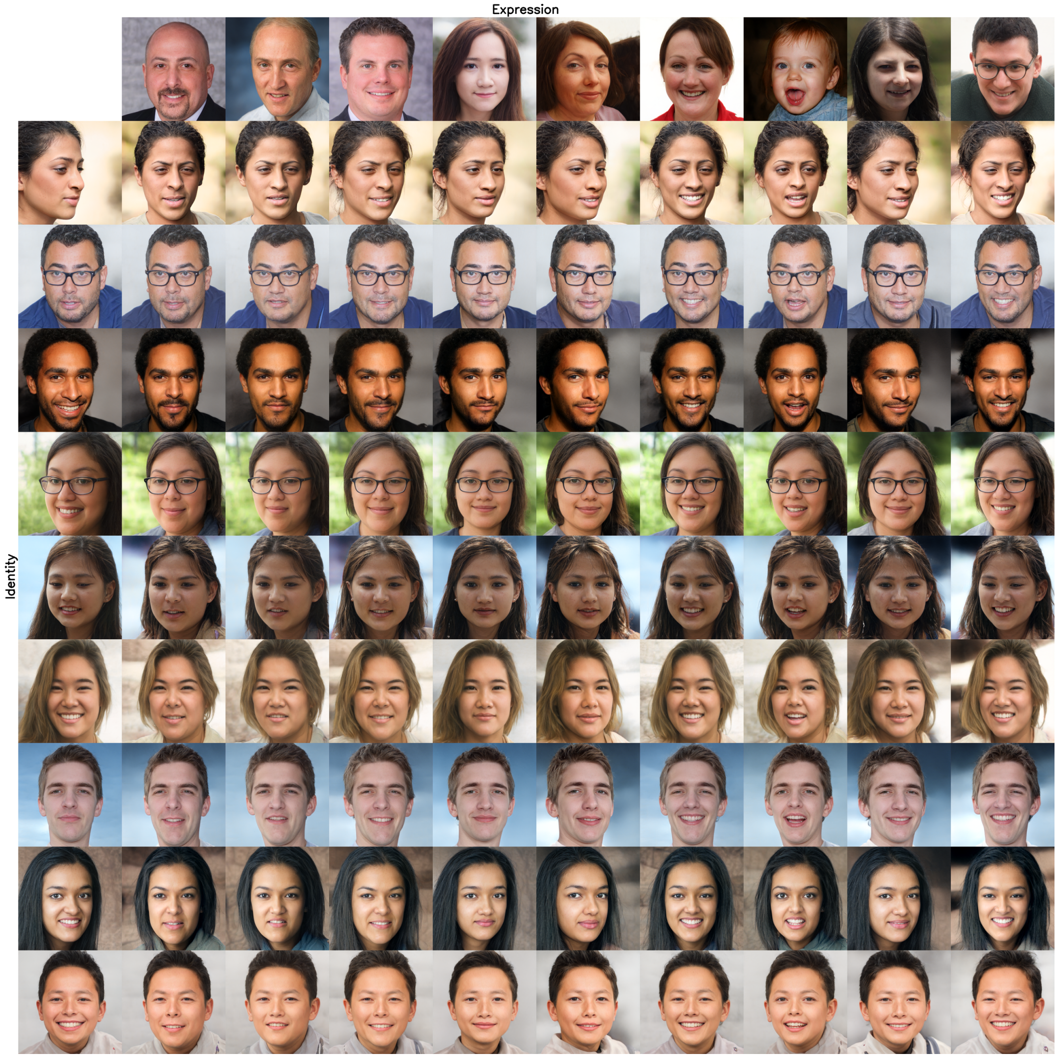

Figure 3: Expression transfer examples of our method. The first row contains the various expression which will be used, and the first column in rows 2-10 contains the identities. We find that the identity is maintained across different expressions.

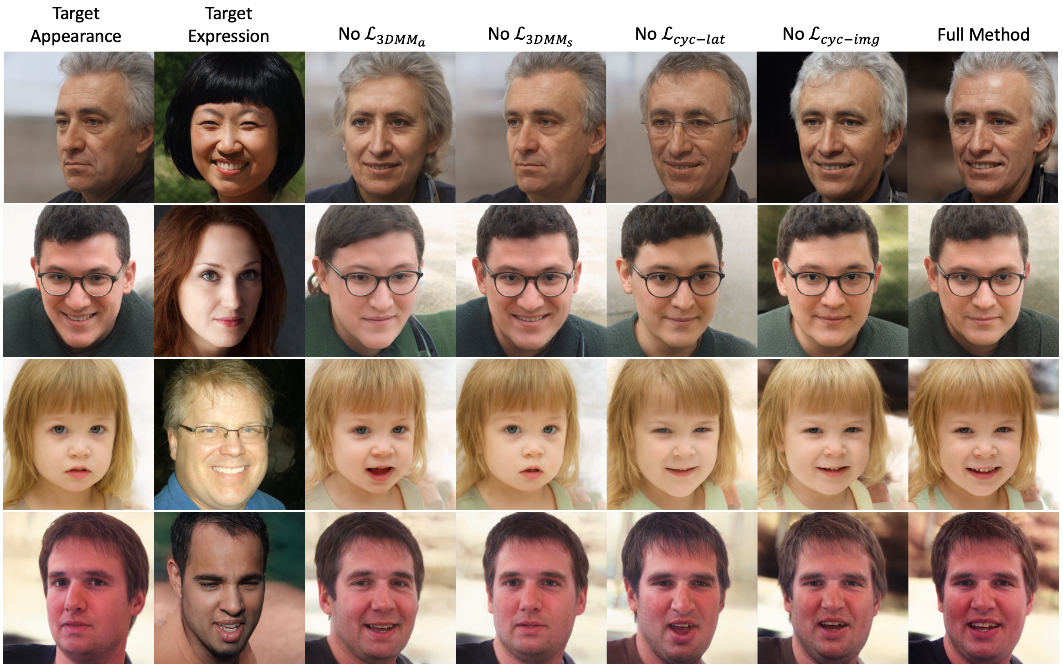
Target Appearance
Target Expression
No $\mathcal{L}_{3DMM_a}$
No $\mathcal{L}_{3DMM_s}$
No $\mathcal{L}_{cyc-lat}$
No $\mathcal{L}_{cyc-img}$
Full Method

Figure 4: Expression transfer examples without auxiliary losses. We find that the full method allows for a balance between preservation of identity and correctly transferring the target expression.

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[2] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[3] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, 2019.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Ankan Bansal, Anirudh Nanduri, Carlos D Castillo, Rajeev Ranjan, and Rama Chellappa. Umdfaces: An annotated face dataset for training deep networks. In *2017 IEEE international joint conference on biometrics (IJCB)*, pages 464–473. IEEE, 2017.

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[7] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

[8] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311. IEEE, 2019.

[9] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pages 5781–5790, 2020.

[10] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPR workshops*, volume 1, 2019.

[11] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.