Controllable and Guided Face Synthesis for Unconstrained Face Recognition — Supplementary Material

Feng Liu, Minchul Kim, Anil Jain, and Xiaoming Liu

Michigan State University, Computer Science & Engineering
{liufeng6,kimminc2,jain,liuxm}@msu.edu

In this supplementary material, we provide:

◇ Additional implementation details including the network structure of the face synthesis model and the training process.

◇ Additional ablation studies, including the effects of different target datasets, the dimensionality of the style coefficient, the perturbation budget, and the ratio of real and synthetic images in each mini-batch.

◇ Additional visualizations of the adversarial training process.

1 Additional Implementation Details

Network Structure. The network architecture of the generator (including the encoder E and decoder G) used in our face synthesis module is illustrated in Tab. 1. We apply Instance Normalization [4] to the encoder and Adaptive Instance Normalization [2] to RESBLOKs (the residual basic block) of the decoder. The encoder takes an image $\mathbf{X}$ with the resolution of 112×112 as input, and outputs its content feature $\mathbf{C} \in \mathbb{R}^{256\times28\times28}$. The input and output to the decoder are $\mathbf{C}$ and the synthesized image $\hat{\mathbf{X}}$, respectively. Additionally, as shown in Fig. 1, the parameters of the Adaptive Instance Normalization (AdaIN) layer in residual blocks are dynamically generated by a multiplayer perceptron (MLP) from the linear subspace model. Following [5], we employ multi-scale discriminators with 3 scales as our discriminator D.

Training Process. We summarize the training process in Tab. 2. In Stage 1, we train our controllable face synthesis module with the identity consistency loss and the adversarial objective. In Stage 2, based on the pre-trained and fixed face synthesis model, we introduce an adversarial regularization strategy to guide the data augmentation process and train the face feature extractor $\mathcal{F}$.

Specifically, in the adversarial FR model training, given B face images $\{\mathbf{X}\}_{i=1}^{B}$ in a mini-batch, our synthesis model (CFSM) is utilized to produces their synthesized version $\hat{\mathbf{X}}$ with initial random style coefficients $\{\mathbf{o}\}_{i=1}^{B}$. Based on the Eqn. 7 and 8 (main paper), we obtain the updated style coefficients $\{\mathbf{o}^*\}_{i=1}^{B}$ with perturbations. We then generate the perturbed images $\{\mathbf{X}^*\}_{i=1}^{B}$ with CFSM. Finally, we randomly select half of $\{\mathbf{X}\}_{i=1}^{B}$ and half of $\{\mathbf{X}^*\}_{i=1}^{B}$ to form a new training batch for the FR model training. Note that, every epoch of the FR model training we will randomly initialize different style coefficients, even for the same training samples.

Table 1. Network architectures of the generator of face synthesis module. RESBLK denotes the residual basic block. [Keys: N=Neurons, K=Kernel size, S=Stride, B=Batch size].

Layer	Encoder (E)	Decoder (G)
1	CONV-(N64,K7,S1), ReLU	RESBLK-(N256,K3,S1)
2	CONV-(N128,K4,S2), ReLU	RESBLK-(N256,K3,S1)
3	CONV-(N256,K4,S2), ReLU	RESBLK-(N256,K3,S1)
4	RESBLK-(N256,K3,S1)	RESBLK-(N256,K3,S1)
5	RESBLK-(N256,K3,S1)	CONV-(N128,K5,S1), ReLU
6	RESBLK-(N256,K3,S1)	CONV-(N64,K5,S1), ReLU
7	RESBLK-(N256,K3,S1)	CONV-(N3,K7,S1), TanH
Output	$\mathbf{C} \in \mathbb{R}^{B\times 256\times 28\times 28}$	$\hat{\mathbf{X}} \in \mathbb{R}^{B\times 3\times W\times H}$

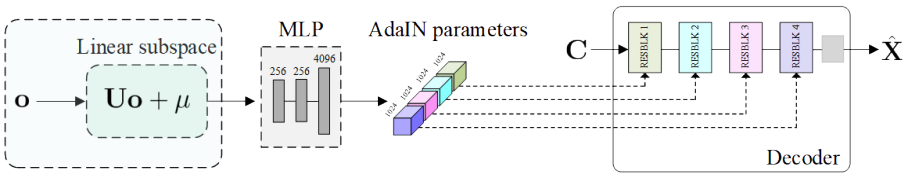

Fig. 1. Additional illustration of the decoder network structure. The parameters of Adaptive Instance Normalization (AdaIN) in residual blocks are dynamically generated by a multiplayer perceptron (MLP) from the linear subspace model.

2 Additional Ablation Studies

Effect of Different Target Datasets. To study how the choice of target dataset in face synthesis model training would affect the face recognition performance, we choose two other datasets, LFW [1] and IJB-S [3] to train the face synthesis models and apply them for the FR model training. During training, for each dataset, we randomly select *unlabeled* 12K face images as the target data to train the face synthesis model. For efficiency, we train the FR models with 0.5M labeled training samples from the MS-Celeb-1M dataset. The diversity of the three face datasets can be ranked as IJB-S > WiderFace > LFW. We show the comparisons on IJB-S protocols in Fig. 2, which shows that the more diverse the unlabeled target dataset is, the more performance gain is obtained. In particular, although LFW is similar to MS-Celeb-1M, it can introduce additional diversity in the dataset when augmented with our controllable and guided face synthesis model. Using *unlabeled* IJB-S images as the target data further improves the performance on the IJB-S dataset, which indicates that our model can be applied for boosting face recognition with limited unlabeled samples available.

Effect of the Dimensionality (q) of the Style Coefficient. Fig. 3 shows the recognition performances on IJB-S over the dimensionality of the style coef-

Table 2. Stages of the training process.

	Network or parameters	Loss
Stage 1	E, G, D, MLP, $\mathbf{U}$, μ	$\mathcal{L}_{ort}$, $\mathcal{L}_{adv}$, $\mathcal{L}_D$, $\mathcal{L}_{id}$
Stage 2	$\mathcal{F}$, δ^*	$\mathcal{L}_{cla}$

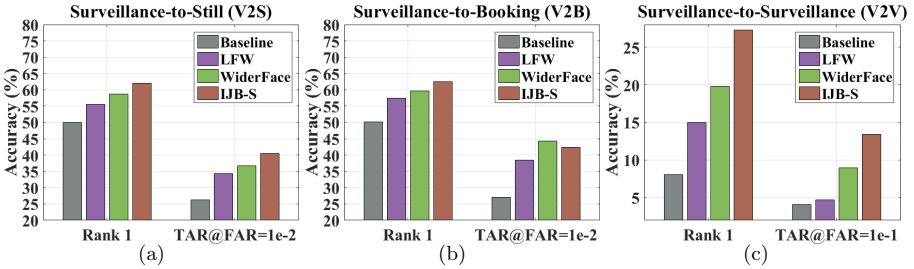

Fig. 2. Evaluation results on IJB-S with different target datasets. **Baseline** refers to the performance of the FR model trained on 0.5 million labeled samples (a subset of MS-Celeb-1M) without using the proposed face synthesis model. In this experiment, other 3 FR models are trained on the 0.5 million labeled samples with the proposed face synthesis models, which are trained with additional 12K unlabeled samples (from LFW, WiderFace or IJB-S, respectively).

ficient. Fig. 3 shows that the dimensionality of the style coefficient does have significant effects on the recognition performance. The model with $q = 10$ performs slightly better in face verification setting, such as V2S and V2B (TAR@FAR=1e-2). The results also indicate that learning manipulation in the low-dimensional subspace is effective and robust for face recognition.

Effect of the Perturbation Budget (ϵ). We conduct experiments to demonstrate the effect of the perturbation budget ϵ. As shown in Fig. 4, we can clearly find that a large perturbation budget ($\epsilon = 0.628$) leads to a better performance in the protocol of Surveillance-to-Surveillance (V2V) while performs slightly worse in the protocols of Surveillance-to-Still (V2S) and Surveillance-to-Booking (V2B). These observations are not surprising because the large style coefficient perturbation would generate faces with low qualities, which is beneficial for improving generalization to the unconstrained testing scenarios.

Effect of the Ratio of Real and Synthetic Images in Each Mini-batch. As illustrated in Sec. 3.2 (main paper), we combine the original real images and their corresponding synthesized version as a mini-batch for the FR model training. In this experiment, we further study the ratio of real (R) and synthetic (S) images in each mini-batch. As shown in Fig. 5, with more synthetic images in each mini-batch (R:S = 25% : 75%), the model achieves the best performance in the most challenging Surveillance-to-Surveillance (V2V) protocol (Rank1).

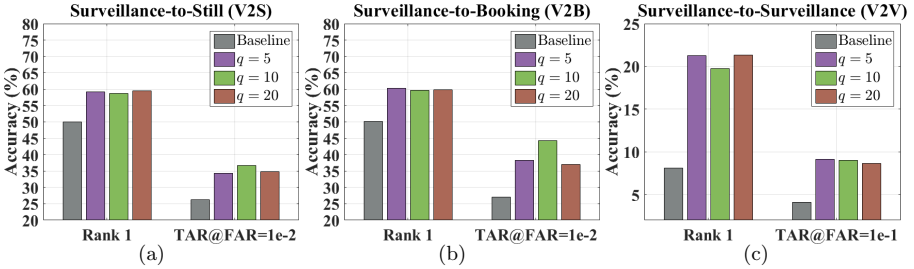

Fig. 3. Evaluation results on IJB-S with the different dimensionalities of the style coefficient ($q = 5$, 10, 20). **Baseline** refers to the performance of the FR model trained on 0.5 million labeled samples (a subset of MS-Celeb-1M) without using the proposed face synthesis model. In this experiment, other 3 models are trained on the 0.5 million labeled samples with the proposed face synthesis model, which is trained with additional 70K unlabeled samples from WiderFace.

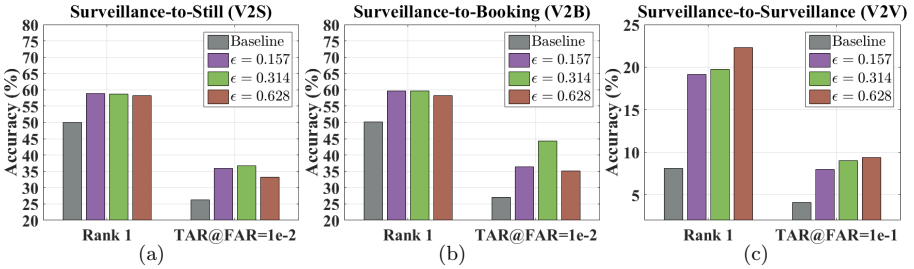

Fig. 4. Evaluation results on IJB-S with different perturbation budget values ($\epsilon = 0.157$, 0.314 or 0.628). **Baseline** refers to the performance of the FR model trained on 0.5 million labeled samples (a subset of MS-Celeb-1M) without using the proposed face synthesis model. In this experiment, other 3 models are trained on the 0.5 million labeled samples with the proposed face synthesis model, which is trained with additional 70K unlabeled samples from WiderFace.

3 Additional Visualizations

The Perturbations in Direction or Magnitude. In adversarial FR model training, our synthesis model is able to offer two meaningful possibilities to perform style coefficient perturbation: magnitude and direction. To study the perturbation properties (direction or magnitude), we collect the initial style coefficient $\mathbf{o}$ and style perturbation $\boldsymbol{\delta}^*$ of 10K samples during the FR model training. We first measure the Cosine Similarity S_C (Fig. 6 (a)) between the initial style coefficient $\mathbf{o}$ and the updated one $\mathbf{o}^* = \mathbf{o} + \boldsymbol{\delta}^*$. Then we present the histogram of the differences (Fig. 6 (b)) between the magnitude of $\mathbf{o}$ and $\mathbf{o}^*$: $a^* - a$, where $a^* = ||\mathbf{o}||$, $a = ||\mathbf{o}^*||$. Finally, in Fig. 6 (c), we show the Sc over $(a^* - a)$. As observed in Fig. 6, the style coefficient perturbation guided by FR model training indeed leads to the changes of both magnitude and direction of the initial style coefficient, which supports the motivation of our controllable

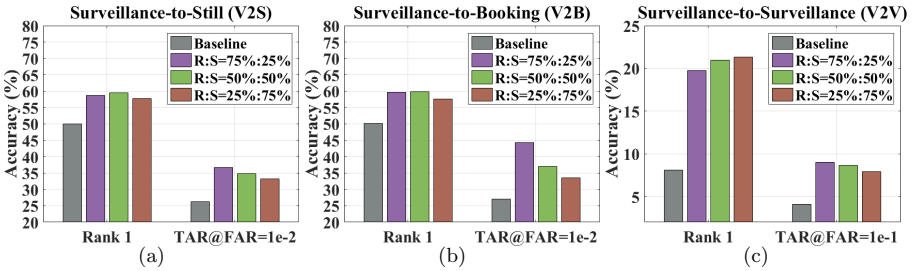

Fig. 5. Evaluation results on IJB-S with different ratios of real (R) and synthetic (S) images in each mini-batch (R:S = 75% : 25%, 50% : 50% or 25% : 75%). **Baseline** refers to the performance of the FR model trained on 0.5 million labeled samples (a subset of MS-Celeb-1M) without using the proposed face synthesis model. In this experiment, other 3 models are trained on the 0.5 million labeled samples with the proposed face synthesis model, which is trained with additional 70K unlabeled samples from WiderFace.

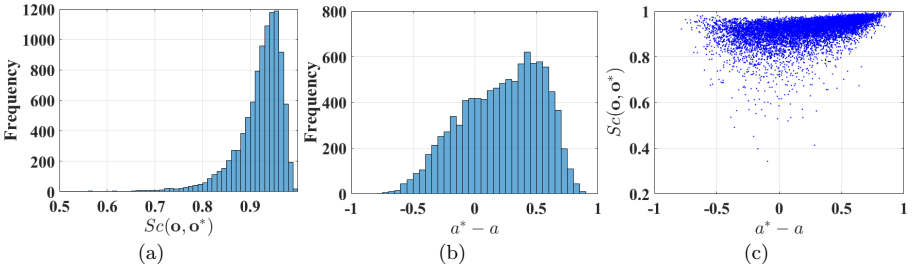

Fig. 6. (a) Histogram of the Cosine Similarity between the initial style coefficient $\mathbf{o}$ and its updated one $\mathbf{o}^*$ with perturbation. (b) Histogram of differences between the magnitude of $\mathbf{o}$ and $\mathbf{o}^*$: $a - a^*$, where $a^* = ||\mathbf{o}||$, $a = ||\mathbf{o}^*||$. (c) Scatter plot showing the correlation between Sc and $a - a^*$.

face synthesis model design. More interestingly, the synthesis model attempts to achieve a balance between magnitude and direction in the adversarial-based augmentation process (see Fig. 6 (c)). For example, when the magnitude is decreasing ($(a - a^*) < 0$), the model is inclined to generate faces in lower quality but more target styles (lower Sc). In contrast, when the magnitude is increasing ($(a - a^*) > 0$), the model prefers to generate faces with higher quality but less target style (larger Sc).

Additional Visualizations of $\mathbf{X}$, $\hat{\mathbf{X}}$ and $\mathbf{X}^*$. In Fig. 7, we show the original examples $\mathbf{X}$, synthesized examples with initial style coefficients $\hat{\mathbf{X}}$ and synthesized examples with style perturbations $\mathbf{X}^*$ in a mini-batch during the FR model training. Additionally, we visualize the pairwise error maps among these 3 types of data. As shown, the guide from the FR model encourages the face synthesis model to generate images with either increased or decreased target face style.

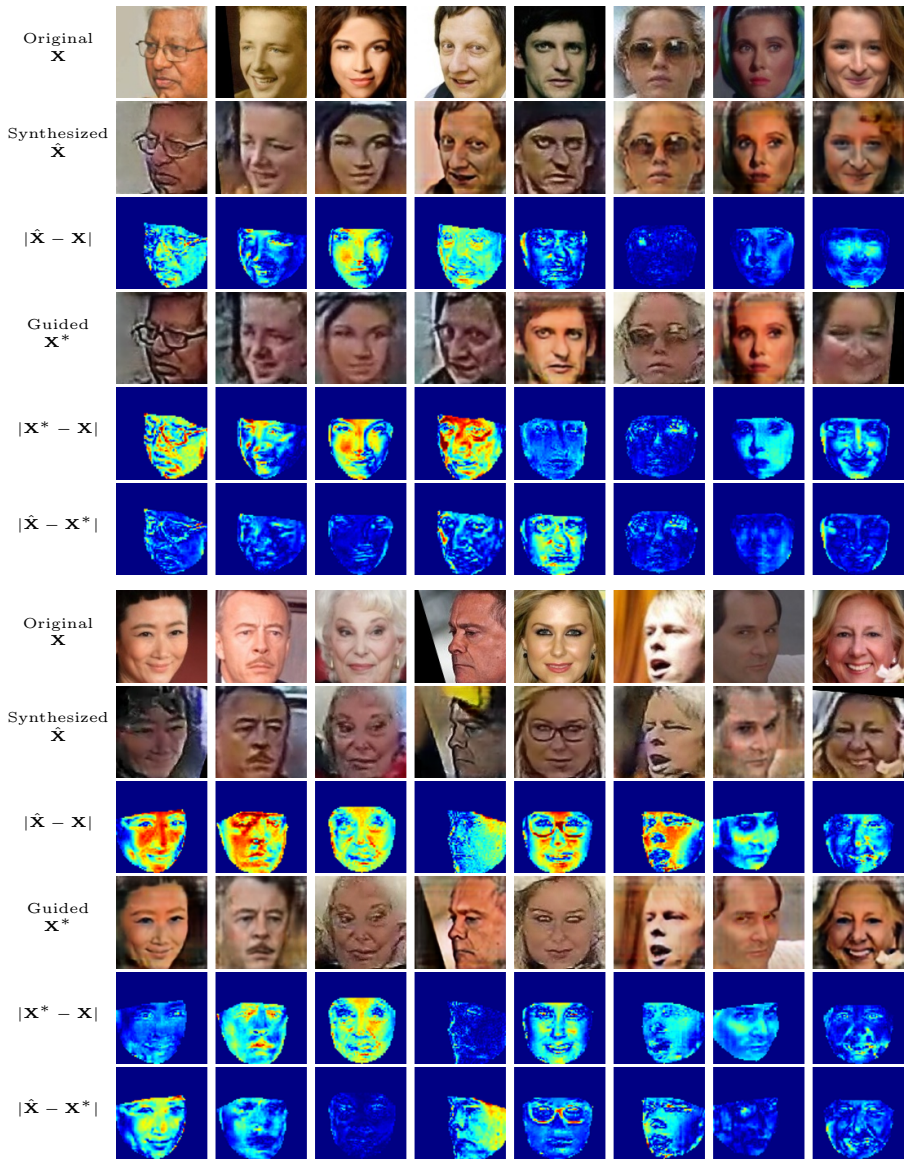

Fig. 7. Training examples in a mini-batch with our face synthesis model during the FR model training. For each image set, we show the original images $\mathbf{X}$, synthesized results with initial style coefficients, and synthesized results with style perturbations $\mathbf{X}^*$. We additionally show their corresponding error maps: $|\hat{\mathbf{X}} - \mathbf{X}|$, $|\mathbf{X}^* - \mathbf{X}|$ and $|\hat{\mathbf{X}} - \mathbf{X}^*|$.

References

1. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (2007) 2
2. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV (2017) 1
3. Kalka, N.D., Maze, B., Duncan, J.A., O'Connor, K., Elliott, S., Hebert, K., Bryan, J., Jain, A.K.: IJB–S: Iarpa janus surveillance video benchmark. In: BTAS (2018) 2
4. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In: CVPR (2017) 1
5. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: CVPR (2018) 1