Few-Shot Head Swapping in the Wild (Supplementary Materials)

Changyong Shu[1] Hemao Wu[2] Hang Zhou[1*] Jiaming Liu[1*] Zhibin Hong[1]
Changxing Ding[2] Junyu Han[1] Jingtuo Liu[1] Errui Ding[1] Jingdong Wang[1]
[1]Department of Computer Vision Technology (VIS), Baidu Inc., [2]South China University of Technology
{zhouhang09,liujiaming03,hongzhibin,hanjunyu,liujingtuo,dingerrui,wangjingdong}@baidu.com,
changyong.shu89@gmail.com, {201830252427@mail.,chxding@}scut.edu.cn.

1. Head2Head Aligner Learning Details

In the training stage of the Head2Head Aligner, the target image I_T and the source image set I_S are all sampled from the same identity. We would expect the output animated portrait I_A of our model to be the same as I_T, thus the whole training procedure is based on frame reconstruction. The loss functions are as follows:

Pixel-wise Reconstruction Loss. The reconstruction loss in the first stage L_{L1}^1 encourages the pixel-wise similarity between the reenactment output I_R and the target image I_T via L1 loss:

$$L_{L1}^1 = \lambda_{L1} \left\| I_T - I_A \right\|_1 . \tag{1}$$

where λ_{rec} is the loss weight.

Perceptual Loss. We utilize the perceptual loss to minimize the semantic discrepancy between the animated portrait I_A and the target image I_T. The feature matching loss is used on a pre-trained VGG-19 network, a pre-trained ResNet with ArcFace [3] E_{id}, and the discriminator D_A:

$$L_{per}^1 = \sum_{k=1}^{K} \sum_{l=1}^{L_k} \lambda_l^k \left\| \phi_l^k(I_T) - \phi_l^k(I_A) \right\|_1 , \tag{2}$$

where λ_l^k balances the terms, ϕ_l^k represents the activation of layer l. $K = 3$ represents the three networks repectively.

Identity Loss. Moreover, to further enhance the ability of identity preservation, we leverage an additional identity loss to minimize the identity gap between the generated output and the target:

$$L_{id} = \lambda_{id}(1 - \cos\left(E_{id}(I_T), E_{id}(I_A)\right)). \tag{3}$$

where λ_{id} is the loss weight, $\cos$ denotes the cosine distance of identity embeddings.

Adversarial Loss. The discriminator D_A is imposed to make the animated portrait I_A looks indistinguishable from the target image I_T. The training losses are defined as follows:

$$L_{adv}^{D_A} = -E[h(D_A(I_T))] - E[h(-D_A(I_A))]. \tag{4}$$

$$L_{adv}^{G_A} = -E[D_A(I_A)], \tag{5}$$

where $h(x) = \min(0, -1 + x)$ is a hinge function used to regularize the discriminator [1, 6].

*Corresponding authors.

2. Head2Scene Blender Details

We jointly train the Semantic-guided Exemplar Warping module and the Blending UNet via the loss functions below, expecting the sequential two jobs to facilitate each other.

Perceptual Loss. Reconstruct training is also leveraged to minimize the difference between the ultimate blending output I_B and the target image I_T by minimizing the perceptual loss.

$$L_{per}^2 = \sum_{l=1}^{L_1} \lambda_l^1 \left\| \phi_l^1(I_B) - \phi_l^1(I_T) \right\|_1 . \tag{6}$$

here λ_l^1 balance the terms layer-wise, ϕ_l^1 represents the activation of layer l in the pre-trained VGG-19 as illustrated in Eq. 2.

Reconstruction Loss. Note that the above feature matching loss excels in capturing fine details, while missing the low frequency image content. This could result in inaccurate colors. Consequently, the L1 loss is also applied for color consistency:

$$L_{L1}^2 = \lambda_1 \left\| I_B - I_T \right\|_1 . \tag{7}$$

Cycle Loss: In order to guarantee that the warped head-color/inpainting exemplars could learn a meaningful correspondence matrix, we introduce the cycle consistent loss.

$$L_c = \lambda_c \left\| I_{T \to A \to T} - I_T \right\|_1 , \tag{8}$$

where $I_{T \to A \to T}$ is the exemplar after cycled warpping, and $I_{T \to A \to T}^k(u) = \sum_{v \in M_A^k} softmax_v(\Gamma^k(u, v)/\tau) \cdot$

$I_{T\to A}(v), u \in M_T^k$. Besides, the additional target image I'_T coming from different image compared to I_A is also utilized to ensure the meaningful of warped exemplar:

$$L_{c'} = \lambda_c \left\| I_{T'\to A\to T'} - I_T \right\|_1 . \tag{9}$$

Adversarial Loss. We utilize another discriminator D_B to distinguish the blending outputs and the real samples from ground truth, with head mask M_A^H and inpainting mask M_A^I concatenated as conditions for further improving the fidelity of our blending outputs. The adversarial objectives are optimized by hinge loss:

$$\begin{aligned} L_{adv}^{D_B} = &-E[h(D_B(I_T \odot M_A^H \odot M_A^I))] \\ &- E[h(-D_B(I_B \odot M_A^H \odot M_A^I))]. \end{aligned} \tag{10}$$

$$L_{adv}^{B} = -E[D_B(I_B \odot M_A^H \odot M_A^I)]. \tag{11}$$

where $\odot$ denotes the concatenation along the dimensionality of channel.

3. Experimental Details

In this section, we describe the data collection and the experimental details, $i.e.$, evaluation metrics, competitors and implementation details.

3.1. Data collection

In terms of Head2Head Aligner, We re-download the 1080P videos with the urls provided by the VoxCeleb2 dataset [8] from YouTube. The frames are aligned by detected landmarks and processed to 512×512. Compared with previous studies [4, 15], a larger cropping window size is used. Under such a setting, we manage to collect overall 28,367 videos with 5,478 different identities, which is much less than the original training set of 145,569 videos with 5,994 different identities in [4, 15]. Besides, totally 805 videos with 86 different identities are gathered as the test set.

3.2. Implementation Details.

The portrait encoder E_{por} is ResNeXt-50 [14], the pose and expression encoder are both constructed by the MobileNetV2 [11]. The size of the latent embedding, $i.e.$ d_1, d_2, d_3 and d_4, are 512, 512, 256 and 256 respectively. The MLP module that transforms them into AdaIN parameters is a 2-layer ReLU perceptron with spectral normalization, where the output of intermediate layer is 768. The generator is borrowed from [4]. We add upsampling residual blocks after the last layer to generate the reenacted image with 512×512 resolution. The first module is trained for roughly three weeks with batch size to be 6 on six 32G Tesla V100 GPUs.

As for the Head2Scene Blender, we use a VGG-19 for feature extraction. The Blending UNet is a seven-layer deep residual U-type network. The training image size is 512×512 and two 48G RTX8000 GPUs is used to train the model for 5 days.

For both models, Adam [9] optimizer is used with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, the imbalanced learning rate for generator and discriminator is set to be 1e-4 and 4e-4 respectively. Spectral normalization is applied to all operators in the system for stabilizing the adversarial training.

4. More Studies on Head2Head Aligner

This section is the supplementary materials for more quantitative and qualitative reenactment comparison under different settings:

One-shot meta-learning evaluation. Fig. 1 is the supplementary qualitative results for cross-id animated portraits in one-shot meta-learning setting. Significant artifacts exhibited in FOMM and Siarohin et al [13]; our method can generate the animated portraits with higher identity similarity and pose consistency, besides, the elaborated emotions (such as happiness in 1st, 4th, and 8th row of Fig. 1) are also well animated.

Impact of K-shot and fine-tuning. Quantitative results of increasing the K-shot number with subject-specific fine-tuning or a meta-learned model are shown in Fig. 2. Qualitative impact of fine-tuning is illustrated in Fig. 3, where better identity preservation is obtained compared to the animated portrait from the meta-learned model, while the pose error is increased for the fine-tuned model overfits the one-shot source. Then we increase the K-shot number, as depicted in Fig. 4, the identity similarity and pose consistency are further improved.

5. More Studies on Head2Scene Blender

This section is the supplementary materials for the discussion of 1) the memory-saving of our semantic-guided warped exemplar module, 2) skin color alignment and 3) inpainting performance in the blending network.

Semantic-guided Exemplar Warping. To verify the mechanism of our proposed calculation-reducing method, we present the semantic-specific contribution that the target image made in the correlation matrix. Specially, the accumulated attention distribution, denoting for pixels in source image with semantic label k, is computed via $\sum_{v\in M_T^k} softmax_v(\Gamma^k(u,v)/\tau)(u \in M_A^k)$, and the qualitative results are exhibited in Fig. 5. Obviously, it demonstrates that the pixels in the target image with different semantic label almost has no contribution in correspondence matrix to that in the source image. Thus the correspondence between pixels from source and target image with different semantic label is redundant. Following our method, roughly averaged 9G GPU memory usage is saved in 512×512 training setting.

Figure 1. Qualitative cross-id animated portraits of one-shot meta-learned model. 1st row: source image for one-shot, 2nd row: pose image from same video but different frame, rows 3 through 6: animated result from FOMM, Siarohin et al [12], LPD and ours. For a vivid show, our animated portraits are re-cropped following LPD.

Skin Color Alignment. For an effective comparison, we overlay the animated portrait on our head swapping result, and using the target head as reference. The qualitative results of skin color alignment with different method on voxceleb2 test dataset is illustrated in Fig. 6, where the remarkable performance advantage on the wild image is depicted. Our method outperforms deepfacelab, SCGAN and PSGAN with more similar skin color to target image, while the haircolor is also aligned and the background remains unchanged. We attribute the superior to that, the skin color exemplar in our method is directly gained from the weighted contribution of the target image, while deepfacelab directly aligns the averaged statistic of the source image to that of the target image; besides, PSGAN and SCGAN learn the implicit skin color information from the inter-feature correspondence and spatially invariant 1D style-code respectively.

Performance of Inpainting. As the missing pixels are affected by the surroundings, for a fair comparison, we excavate neck and background around the head from our head swapping results and take them as input for inpainting, making facial skin color consistent with the skin color. Since the source code of reference-guided competitors (LOA [16] and TransFill [17]) are not released publicly, we send the testing data to authors for the compared result respectively. Qualitative comparison results are shown in Fig.7.

6. More head swapping results

Additional comparison results with other methods on VoxCeleb2 test set are shown in Fig. 8-22.

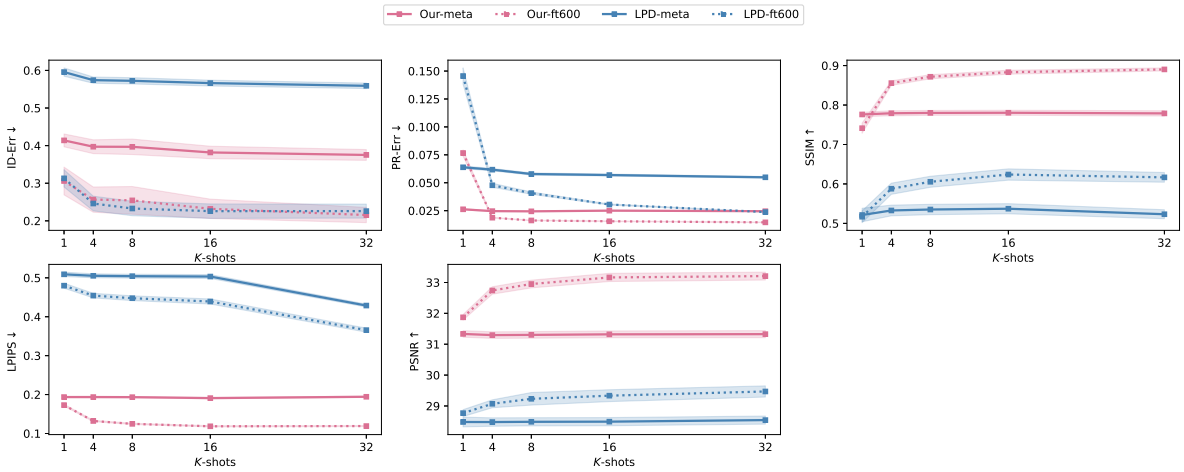

Figure 2. Quantitative result of increasing the K-shot number with subject-specific fine-tuning or a meta-learned model. Meta denotes the meta-learned model, ft600 indicates the fine-tuned model under 600 iterations. The transparent area around the dotted and solid lines represents the variance.

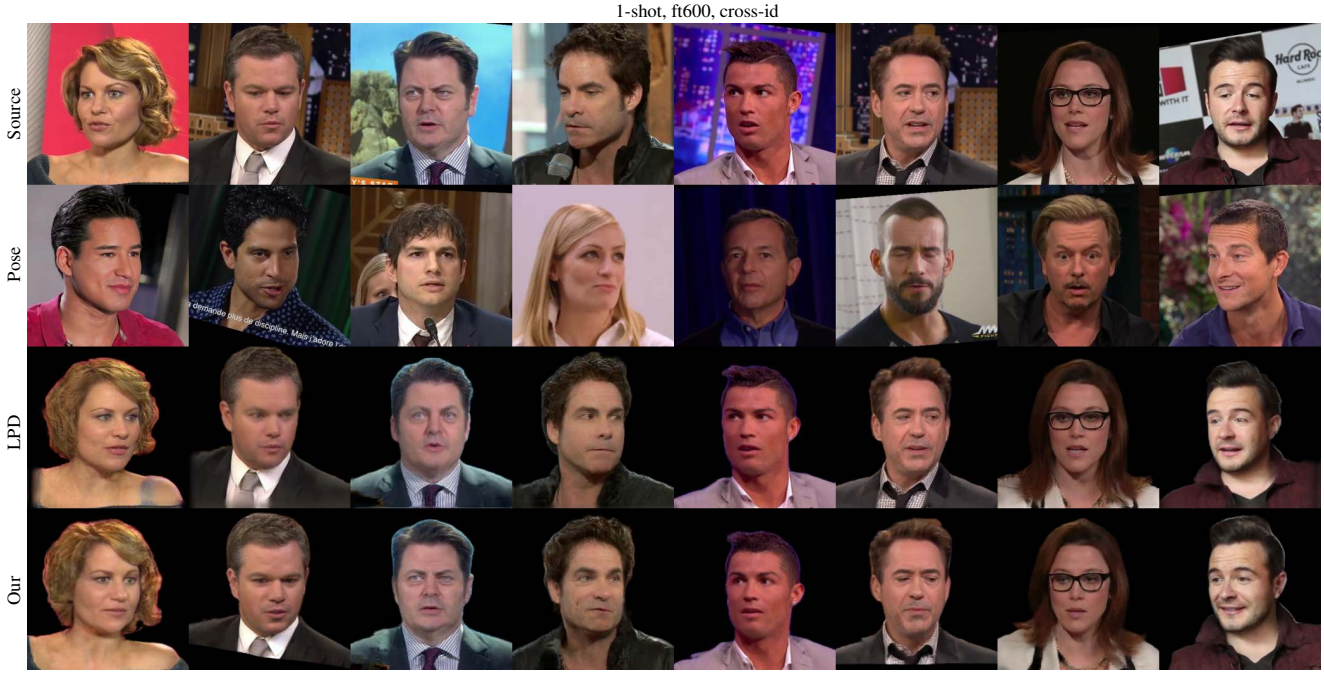

Figure 3. Qualitative cross-id animated portraits of one-shot ft600 model, the layout is the same as in Fig. 1.

Figure 4. Qualitative cross-id animated portraits of 32-shot ft600 model, the layout is the same as in Fig. 1.

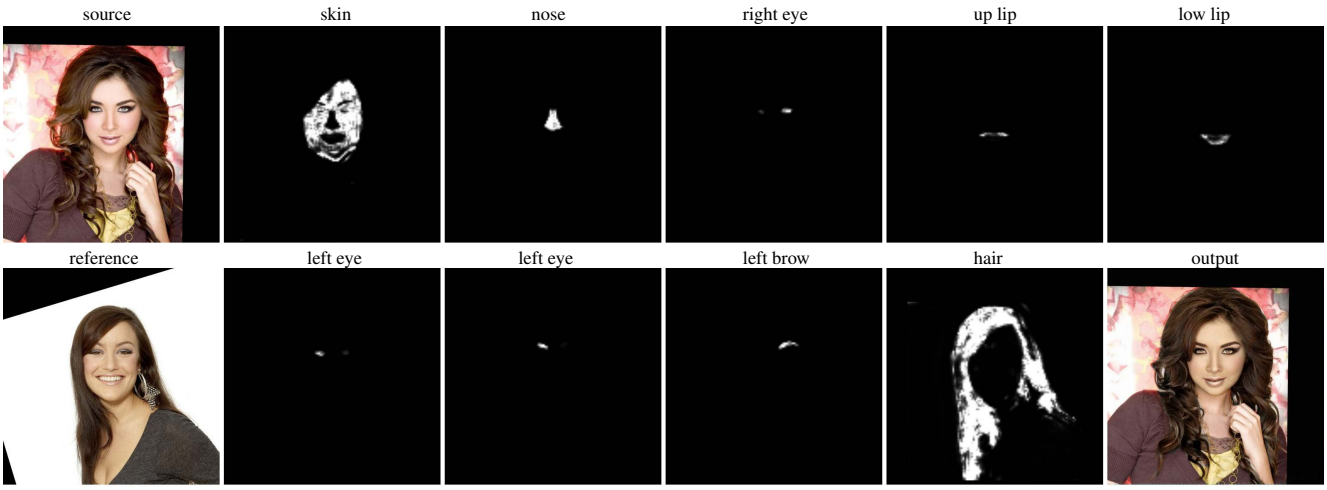

Figure 5. The accumulated attention distribution for diverse semantic label, it shows an intuitive phenomenon: the highlighted regions correspond to semantic label blow the sub figure, and the remain regions are pitch-dark.

Figure 6. Qualitative comparsion of facial skin color alignment with different method on voxceleb2 test dataset, best viewed in color.

Figure 7. Reference-guided inpainting comparison.

Figure 8. Additional comparison results with other methods on VoxCeleb2 test set. The top left image is one of the 32 source images. The remaining images in the 1st row are the pose images. The 2nd line plots the state-of-the-art face swapping results [2] by replacing the face in target image with that in source reenactment. The 3rd line shows the head swapping result of deepfacelab [10]. And the last line shows our head swapping results. Best viewed in color.

Figure 9. Additional comparison results with other methods on VoxCeleb2 test set, the layout is the same as in Fig. 8.

Figure 10. Additional comparison results with other methods on VoxCeleb2 test set, the layout is the same as in Fig. 8.

Figure 11. Additional comparison results with other methods on VoxCeleb2 test set, the layout is the same as in Fig. 8.

Figure 12. Additional comparison results with other methods on VoxCeleb2 test set, the layout is the same as in Fig. 8.

Figure 13. Additional comparison results with other methods on VoxCeleb2 test set, the layout is the same as in Fig. 8.

Figure 14. Additional comparison results with other methods on VoxCeleb2 test set, the layout is the same as in Fig. 8.

Figure 15. Additional comparison results with other methods on VoxCeleb2 test set, the layout is the same as in Fig. 8.

Figure 16. Additional comparison results with other methods on VoxCeleb2 test set, the layout is the same as in Fig. 8.

Figure 17. Additional comparison results with other methods on VoxCeleb2 test set, the layout is the same as in Fig. 8.

Figure 18. Additional comparison results with other methods on VoxCeleb2 test set, the layout is the same as in Fig. 8.

Figure 19. Additional comparison results with other methods on VoxCeleb2 test set, the layout is the same as in Fig. 8.

Figure 20. Additional comparison results with other methods on VoxCeleb2 test set, the layout is the same as in Fig. 8.

Figure 21. Additional comparison results with other methods on VoxCeleb2 test set, the layout is the same as in Fig. 8.

Figure 22. Additional comparison results with other methods on VoxCeleb2 test set, the layout is the same as in Fig. 8.

References

[1] Brock Andrew, Donahue Jeff, and Simonyan Karen. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1

[2] Renwang Chen, Xuanhong Chen, and Yanhao. Ge. Simswap: An efficient framework for high fidelity face swapping. In *ACM Int. Conf. Multimedia*, 2020. 7, 8, 9, 10, 11, 12, 13, 14

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1

[4] Burkov Egor, Pasechnik Igor, Grigorev Artur, and Lempitsky Victor. Neural head reenactment with latent pose descriptors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[5] Han, Chu Deng, Hongmin Han, Guoqiang cai, Shengfeng Han, and He. Spatially-invariant style-codes controlled makeup transfer. In *CVPR*, 2021. 6

[6] Zhang Han, Goodfellow Ian, Metaxas Dimitris, and Odena Augustus. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018. 1

[7] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. X2face: A network for controlling face generation by using images, audio, and pose codes. *cvpr*, 2020. 6

[8] Son Chung Joon, Nagrani Arsha, and Zisserman. Andrew. Voxceleb2: Deepspeaker recognition. In *IEEE Conf. Conference of the International Speech Communication Association.*, 2018. 2

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[10] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Um, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, et al. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*, 2020. 6, 7, 8, 9, 10, 11, 12, 13, 14

[11] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2018. 2

[12] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. *CVPR*, 2019. 3

[13] Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021. 2

[14] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, July 2017. 2

[15] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Int. Conf. Comput. Vis.*, October 2019. 2

[16] Tong Zhou, Changxing Ding, Shaowen Lin, Xinchao Wang, and Dacheng Tao. Learning oracle attention for high-fidelity face completion. *CVPR*, 2020. 3, 6

[17] Yuqian Zhou, Connelly Barnes, Eli Shechtman, and Sohrab Amirghodsi. Transfill: Reference-guided image inpainting by merging multiple color and spatial transformations. *CVPR*, pages 2266–2276, 2021. 3, 6