Faithful Face Image Completion for HMD Occlusion Removal

Miao Wang[1]* Xin Wen[1] Shi-Min Hu[2]

1 State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China
2 Department of Computer Science and Technology, Tsinghua University, Beijing

Figure 1: Given a facial image (a)(d) partially occluded by HMD and an occlusion-free reference image (b)(e), our method faithfully synthesizes the occluded facial content (c)(f).

ABSTRACT

Head-mounted-displays (HMDs) provide immersive experiences of virtual content. While being flexible, HMDs could be a hindrance for Virtual Reality (VR) applications such as VR teleconference where facial components and expressions of the user are partially occluded thus cannot be seen by others. We present an automatic face image completion solution that treats the occluded region as a hole and completes the hole with the help of an occlusion-free reference image of the same person. Given the occluded input image and an occlusion-free reference image, our method first computes head pose features from estimated facial landmarks. The head pose features, as well as images, are then fed into a generative adversarial network (GAN) to synthesize the output image. Our method can generate faithful results from various input cases and outperforms other face completion methods. It provides a light-weighted solution to HMD occlusion removal and has the potential to benefit VR applications.

Index Terms: Computing methodologies—Computer graphics—Image manipulation—Image processing;

1 INTRODUCTION

With the rapid development of visual computing and hardware technology, various types of portable head-mounted displays (HMDs) such as Facebook Oculus, HTC Vive and Microsoft HoloLens have been developed. The integration of HMD and other sensors (e.g. controller) provides immersive and high-quality experiences of virtual content. In some Virtual Reality (VR) applications such as VR teleconference, VR education and VR entertainment, the post-processing of HMD removal in images and videos are essential for improving communication experience. In this paper, we propose to complete the missing facial region occluded by HMD using an image synthesis solution. The solution is purely algorithmic without any hardware requirements. Our method only requires a reference image of the user without any occlusion to faithfully recover the identity. Given an occluded image and a reference image, our method synthesizes an output image with the occluded region filled with faithful content consistent with the head pose.

The challenge in this problem is that even though a reference image of the same person is provided, the head poses, illuminations, tones and backgrounds of the input and reference images could be different. How to fill the large hole with faithful content is challenging. To overcome the challenge, we propose to synthesize the facial content using a generative adversarial network (GAN). The network takes the occluded input image and reference image, as well as the head pose features computed from facial landmarks as inputs, and outputs an occlusion-completed image. The result is identity-preserved with a natural appearance.

Compared with many prior generative face image completion works [13, 17, 18], our method is able to preserve identity of the user. While the recent work [25] proposed an identity-preserving face completion method to this problem, the result could be blurry. In our method, the head poses of input and reference face images are introduced as conditions. A face completion network consists of one generator and two pairs of discriminators are proposed to complete the occluded region. Our method can benefit communications in VR teleconference, education and entertainment, etc.

The contributions of this work are: 1) we present a deep learning-based faithful face image completion method for HMD occlusion removal, it is light-weighted without relying on hardwares and provides new insights for headset removal in VR applications; 2) we present head pose features and pose consistency discriminators to improve sharpness and preserve the consistency between the completed content and the head pose, which are demonstrated as key components of the method.

2 RELATED WORK

Our work is related to face modeling solutions typically assisted with additional capture hardwares. As we propose an image processing solution, our work is also closely related to image synthesis works.

*e-mail:miaow@buaa.edu.cn

978-1-7281-4765-9/19/$31.00 ©2019 IEEE
DOI 10.1109/ISMAR-Adjunct.2019.00-36

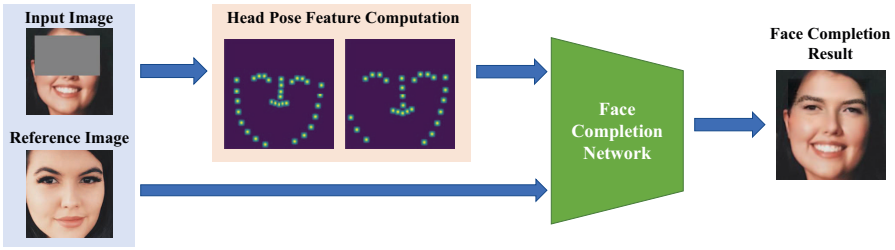

Figure 2: The pipeline of our method. Given an occluded input image and an occlusion-free reference image, we compute head pose features from the images. The features as well as input images are sent into a face completion network to generate the completed face image.

2.1 Face Modeling

Recent advancements in computer graphics and computer vision have promoted the development of 3D facial performance capture methods using single RGB camera [2]. Face reenactment methods [11, 21] were introduced to transfer facial expressions or the whole portrait from one person to another in real time. Face reenactment can obtain high-quality 3D-based expression reconstruction, but it cannot work when the user is equipped with headsets. Takemura and Ohta [19] developed a face synthesis system that uses an infrared position sensor with an image-based rendering algorithm to diminish HMD for shared mixed reality. FaceVR system [20] was proposed to reconstruct the user's face model when she is mounted with HMD, with two infrared cameras inside the HMD to capture eye content and one RGB-D camera outside the HMD to capture the whole face. A similar setup was proposed by Chen et al. [3] to obtain the facial expression to drive avatar motion. The above solutions require infrared cameras fixed inside the headset to faithfully reconstruct the user's identity and eye content. Different from these works, we propose a light-weighted image processing solution.

2.2 Face Image Editing with GANs

Goodfellow et al. [7] first introduced the concept of the generative adversarial networks (GANs), consisting of a generator and a discriminator. During the training, the generator and discriminator compete against each other to improve their network ability for better content generation and discrimination. Ever since its introduction, the GAN method has been widely applied to image-to-image translation and image completion tasks. Isola et al. [10] proposed a GAN network that "translates" an image to another domain, such as from sketch to photo, from architectural maps to photo, from black-and-white to color photos, etc. In image completion task [15], the contents of an arbitrary image region conditioned on its surroundings are generated by a convolutional neural network. Later, Iizuka et al. [9] proposed an image completion network with global and local discriminators. The novel local discriminator helps scrutinize the details of the completed image.

GAN has also been widely used in face image editing applications such as face completion [13, 18, 22, 25], face swapping [12, 14], sketched-based editing [17], etc. In the face completion problem, Li et al. [13] proposed a network built upon [9] with a novel face parsing module. Song et al. [18] developed a two-phase network which first estimate the facial geometry and then use the image and geometry information to complete the missing region. Korshunova et al. [12] developed an algorithm for swapping faces, trained on dozens of images of an individual person. On the contrary, our technique presents a general network ready for the test on new person. Closely related to our method, Zhao et al. [25] proposed the state-of-the-art method to complete the missing region occluded by HMD while preserving the identity. In their method, an occluded input image, an occlusion-free reference image of the same person and a target pose map of the generated face are given, the network synthesizes the missing content with the identity preserved. While it achieved appealing results, it requires a pre-determined target pose map while not guaranteeing the consistency between synthesized content and head pose. Our method is capable of synthesizing faithful results consistent with the head pose.

3 Method

Formulating the HMD occlusion removal task as a face image completion problem, we propose to solve it using a deep convolutional neural network. To complete the face image x with an occluded region o, we assume that an occlusion-free reference image of the same person r is also given. We first estimate the head pose features h_x and h_r from the input image x and the reference image r respectively. The head pose features are essential for faithful face completion. After that, we feed the input and reference images into a completion network conditioned on the estimated head pose features h_x and h_r, to generate an image y' with the missing region completed. The overall pipeline of our method is illustrated in Figure 2.

3.1 Head Pose Feature Computation

Head pose feature computation is a key component of our method. It is used to improve the visual consistency between the completed content and the head pose. To compute the head pose features, we estimate the facial landmarks from the input image x and the reference image r. We use the off-the-shelf face alignment algorithm [1] as a basic technique to compute the 68-point facial landmarks L_x or L_r from the input image x or reference image r. However, because the input image x has a large occluded region, even though the estimated landmarks L_x are reasonable, they may not reveal the identity of the face. To improve the accuracy of landmark estimation in the input image, we propose to warp the subset L_r^o of landmarks L_r that corresponds to the landmarks $L_x^o \subset L_x$ in the occluded region o in image x, with a transformation H:

$$\widehat{L}_r^o = H \cdot L_r^o, \tag{1}$$

where $\widehat{L}_r^o$ is the warped landmark set of L_r^o, H is the Homography transformation matrix estimated from facial landmarks $(L_r - L_r^o)$ to $(L_x - L_x^o)$ using RANSAC algorithm [6]. This landmark refinement process is only performed when the warping is robust. In our implementation, if the yaw angle of the head is smaller than 15° (see Figure 3 (a)), the occluded landmarks L_x^o will be substituted by the warped landmarks $\widehat{L}_r^o$.

The head pose features encode facial landmarks which are consistent with the head pose while not sensitive to facial expressions. They include five components: the face contour, the bridge of the nose, the tip of the nose, the left upper eyelid and the right upper eyelid; each component is encoded into a one-channel feature map. We assign a Gaussian kernel at each landmark position with mean value $\mu = 0.0$ and standard deviation $\sigma = 2.0$ in corresponding feature map channel (visualized in Figure 3 (b)). The five channels corresponding to five facial components are concatenated as the head pose feature. We experimented with encoding more components,

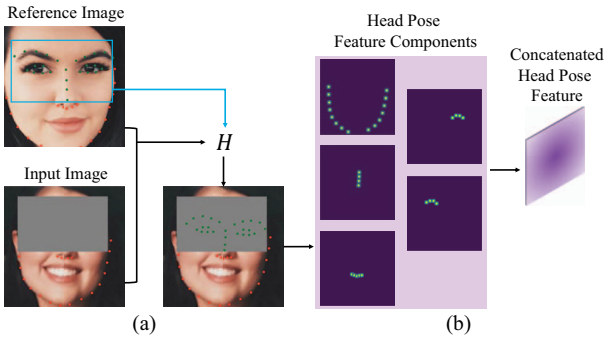

Figure 3: Head pose feature computation. (a) shows the facial landmark refinement of L_x^o by warping landmarks L_r^o (marked in green color) using a Homography transformation H. The transformation is computed from $L_r - L_r^o$ and $L_x - L_x^o$ (marked in red color); (b) shows the five robust landmark components which are encoded into the head pose feature.

Table 1: Architecture of the generator. After each convolutional layer, except the last one, there is a Exponential Linear Unit (ELU) layer. The output layer consists of a convolutional layer with a sigmoid function instead of a ELU layer. "Outputs" refers to the number of output channels for the output of the layer.

Type	Kernel	Dilation	Stride	Outputs
Conv.	5×5	1	1×1	64
Conv.	3×3	1	2×2	128
Conv.	3×3	1	1×1	128
Conv.	3×3	1	2×2	256
Conv.	3×3	1	1×1	256
Conv.	3×3	1	1×1	256
Dilated Conv.	3×3	2	1×1	256
Dilated Conv.	3×3	4	1×1	256
Dilated Conv.	3×3	8	1×1	256
Dilated Conv.	3×3	16	1×1	256
Conv.	3×3	1	1×1	256
Deconv.	4×4	1	1/2×1/2	128
Conv.	3×3	1	1×1	128
Deconv.	4×4	1	1/2×1/2	64
Conv.	3×3	1	1×1	32
Conv	3×3	1	1×1	3

however we hardly observed improvement on quality while reducing efficiency. The pose features h_x and h_r together with the face images are fed into the face completion network.

3.2 Face Completion Network

The face completion network learns to complete the missing region with content consistent with the head pose while maintaining the identity of the face. We adapt the image synthesis network from [9, 24] which uses global and local discriminators in a GAN architecture to achieve global and local sense of compatibility between completed content and original content. In our face completion problem, in order to preserve the identity information and obtain the pose compatibility between the completed result y' and input x, we propose to use a conditional GAN architecture with additional extracted pose features as conditions.

The network architecture is illustrated in Figure 4 and listed in Table 1 and Table 2. The generator G learns to generate an output image $y' = G(x, h_x, r, h_r)$ from inputs $\{x, h_x, r, h_r\}$, while the discriminators try their best to distinguish the generated image from real ones. To ensure that the completed content is photorealistic,

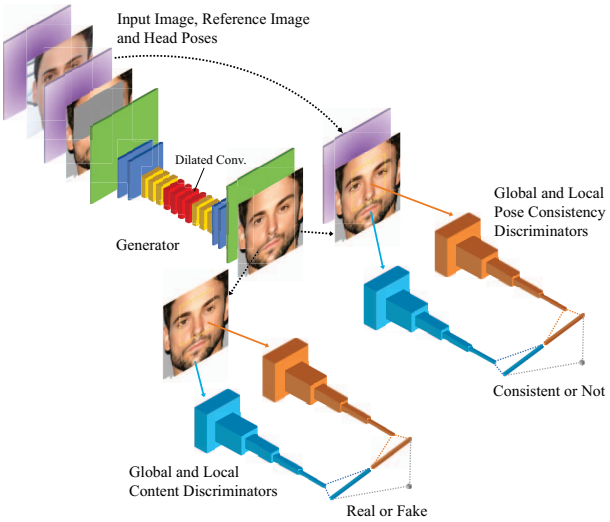

Figure 4: Completion network architecture. It consists of a generator and two pairs of discriminators: pose consistency discriminators and content discriminators. The generator takes the occluded input image, occlusion-free reference image and corresponding pose features as input, and predicts a completed face image. The pose consistency discriminators are trained to determine if an image is consistent with the head pose. The content discriminators are trained to determine if the image content contains a real face.

Table 2: Architectures of the discriminators. All Conv. layers are followed with an ELU activation. Fully-Connected (FC) layers refer to the standard neural network layers. The output layer consists of a fully-connected layer with a sigmoid transfer layer that outputs the probability that an input image came from real images rather than the completion network.

(a) Global and Local discriminator

Type	Kernel	Stride	Outputs
Conv.	5×5	2×2	64
Conv.	5×5	2×2	128
Conv.	5×5	2×2	256
Conv.	5×5	2×2	512
Conv.	5×5	2×2	512
FC	-	-	1024

(b) Concatenation layer

Type	Kernel	Stride	Outputs
Concat.	-	-	2048
FC	-	-	1

a pair of content discriminators $D_C = \langle D_C^G, D_C^L \rangle$ is employed. The content discriminators are a global discriminator D_C^G and a local discriminator D_C^L. The global discriminator D_C^G measures the output from a global perspective, while the local discriminator D_C^L takes a small patch $M(y')$ of the output y' and determines whether it is consistent with the surroundings.

The adversarial objective associated with G and D_C is given as:

$$\begin{aligned}\mathcal{L}_{GAN}^C(G, D_C) = {} & \mathbb{E}_y[\log D_C^G(y)] + \mathbb{E}_y[\log D_C^L(M(y))] \\ & + \mathbb{E}_{x,h_x,r,h_r}[\log(1 - D_C^G(h_x, G(x, h_x, r, h_r)))] \\ & + \mathbb{E}_{x,h_x,r,h_r}[\log(1 - D_C^L(h_x, M(G(x, h_x, r, h_r))))],\end{aligned} \tag{2}$$

where G tries to minimize it against the adversarial D_C that tries to maximize it.

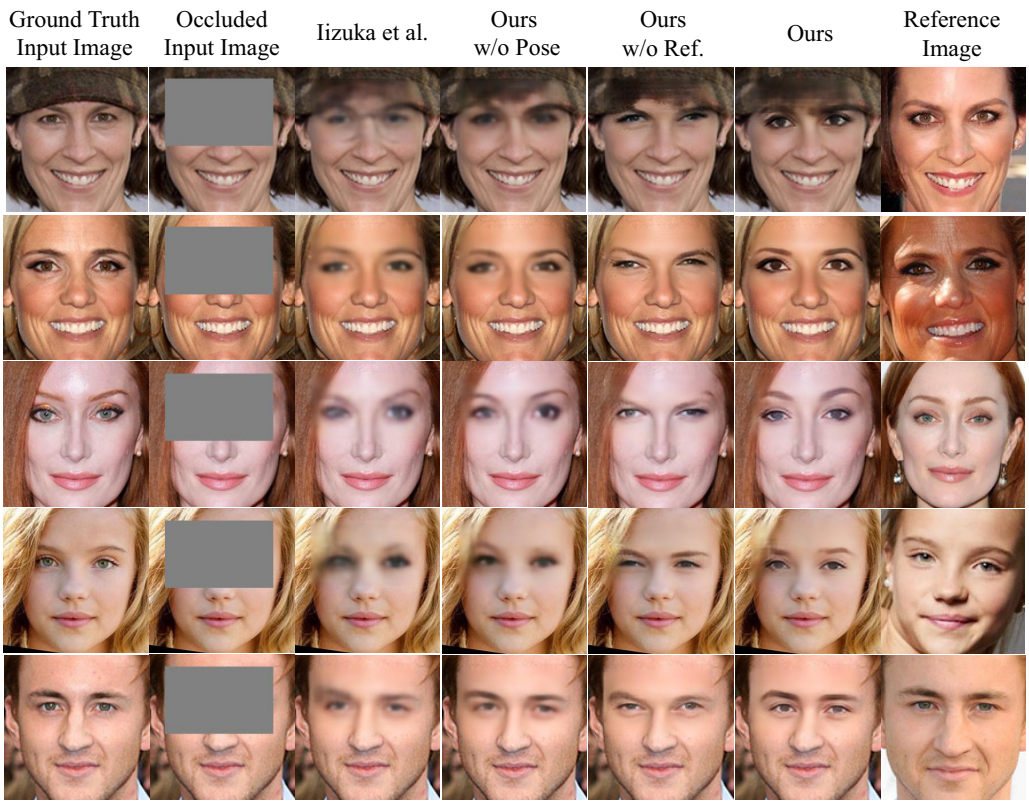

Figure 5: Qualitative comparison with alternative methods.

With G and D_C, the network is able to synthesize facial content, but the result could be blurry or not consistent with the head pose. To avoid this, we present pose consistency discriminators $D_P = \langle D_P^G, D_P^L \rangle$ to encourage the synthesized content aligned with the extracted head pose. The structure of the pose consistency discriminators resembles the content discriminators with additional conditional inputs: D_P^G takes the completed image y' together with the pose feature h_x as inputs, while D_L^G takes a patch $M(h_x, y')$ of them as inputs. The pose consistency discriminators are trained to determine whether the content inside the completed region is compatible with the head pose. The adversarial objective associated with the pose consistency discriminators is defined as:

$$\begin{aligned}\mathscr{L}_{GAN}^{P}(G,D_P) = \mathbb{E}_{h_x,y}[\log D_P^G(h_x,y)] + \mathbb{E}_{h_x,y}[\log D_P^L(M(h_x,y))] \\ + \mathbb{E}_{x,h_x,r,h_r}[\log(1 - D_P^G(h_x, G(x,h_x,r,h_r)))] \\ + \mathbb{E}_{x,h_x,r,h_r}[\log(1 - D_P^L(M(h_x, G(x,h_x,r,h_r))))]\end{aligned} \tag{3}$$

To faithfully reconstruct the face component and further preserve identity, we introduce an L1 loss between $G(x)$ and the ground truth input image y without occlusion:

$$\mathscr{L}_{L1}(G) = \mathbb{E}_{x,h_x,r,h_r}[||y - G(x,h_x,r,h_r)||_1], \tag{4}$$

note that we have both the content discriminators and the pose consistency discriminators to ensure face realism and content consistency with head pose, the network is capable of preserving identity as well as producing sharp results using the L1 loss.

Combining the above loss functions, the final objective of the completion network is:

$$\begin{aligned}G^* = \arg\min_G \max_{D_C, D_P} \mathscr{L}_{L1}(G) + \lambda_1 \mathscr{L}_{GAN}^{C}(G, D_C) \\ + \lambda_2 \mathscr{L}_{GAN}^{P}(G, D_P),\end{aligned} \tag{5}$$

where λ_1 and λ_2 are parameters balancing the loss terms. With the above objective function, we can train a network that predicts faithful content within the occluded region. Once the network outputs a completed image y', we perform Poisson blending [16] between the completed content and the remaining un-occluded content to synthesize the final face image.

4 Experiments

4.1 Implementation Details and Run-time Performance

During training, the occluded region is determined using the bounding box of eye and nose landmarks. In the completion network, the input image size of generator is 256×256, and the input size of local discriminator is 128×128. The objective is optimized using Adadelta [23] with default parameters. The batch size is set as 12 and the weights λ_1 and λ_2 in the objective are set as 0.0004. ELU activation function [4] is used after all convolution layers. The training is composed of three phases as in [9]; it took about 3 days on a single NVIDIA GTX 1080 Ti graphic card. The inference process of the network takes about 0.013 second on the same graphic card, which supports real-time applications.

4.2 Dataset

The network is trained on Celeb-ID dataset [5]. This dataset is originally built for eye region in-painting of images, with face regions cropped and resized to 256×256 in advance. It contains about

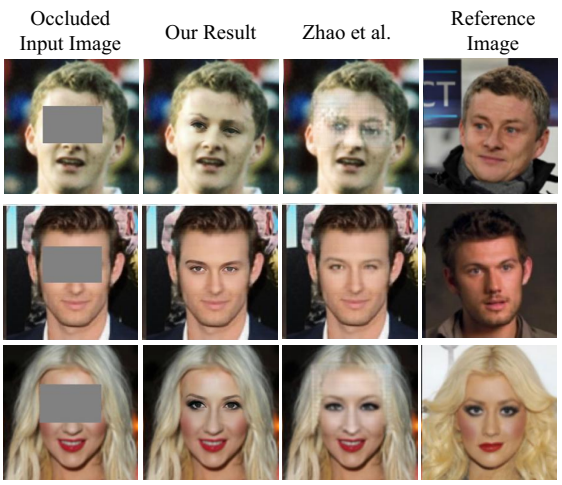

Figure 6: Comparison with the state-of-the-art face completion method.

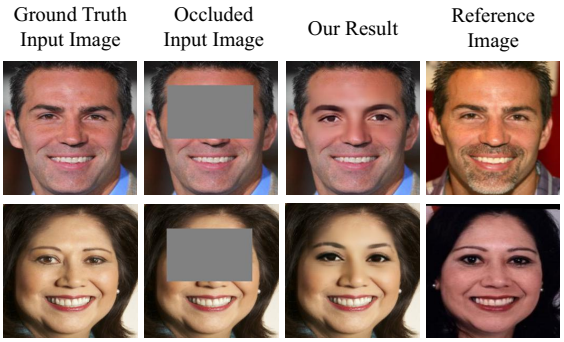

Figure 7: In the wild tests using MS-Celeb-1M data [8]. The tones from input images and reference images are different.

17K individuals with at least 3 portrait photographs of each person. Blurry and monochrome images, as well as images with extreme head poses are filtered out manually. As a result, 78K pairs of images are collected to train the completion network and 700 pairs are for testing; training and testing data have no identity overlap. With the pre-trained model using Celeb-ID dataset, we also test our method on MS-Celeb-1M face images [8] which was originally designed for face recognition (see Figure 7). Portrait video data collected from identity-preserving face completion [25] are tested for video completion application (see Figure 8 and Figure 9).

4.3 Baseline Methods

We compare our method with globally and lobally consistent image completion [9] (denoted as *Iizuka et al.*), identity preserving face completion [25] (denoted as *Zhao et al.*) and our alternative configurations without pose consistency adversarial loss $\mathscr{L}_{GAN}^{P}$ and conditional pose features (denoted as *Ours w/o Pose*), and ours without using the reference image (denoted as *Ours w/o Ref.*).

4.4 Results

We compare our face completion method with the state-of-the-art methods. As shown in Figure 5, while Iizuka et al. shown sharper results in their paper [9], the completion results for large occluded regions are still blurry. To demonstrate the power of the proposed pose consistency discriminators in our method, we also compared our method with *Ours w/o Pose*. It can be seen that results from *Ours w/o Pose* are clearer than Iizuka et al.'s results, but are inferior to our final results. Without using the reference image, the results from *Ours w/o Ref.* do not preserve the identity well. We compare with *Zhao et al.* on their test set. Figure 6 shows visual comparisons between our results and *Zhao et al.*. Although their method generally synthesizes reasonable results, some completed faces could be blurry if the input and reference are not close enough in head pose, illumination or tone. Moreover, their method requires manually determined pose as input. Our method is fully automatic and produces visually sharper and clearer results.

In the wild tests. We test our model using images collected from a different dataset MS-Celeb-1M [8]. As shown in Figure 7, although the tones of the occluded image and the reference image are different, our method is able to synthesize appealing results.

Test on videos. We further test our model on video frames. To perform face completion, we first crop the face region from each original frame based on facial landmark detection. After that, we complete the occluded region using the proposed method. Finally, we paste the face region back to the original frame. The steps are performed frame by frame through the timeline. Figure 8 shows the completed frames with different head poses from a video sequence using only one reference frame. Regardless of the variations of poses, our face completion network can produce visually plausible results. Figure 9 shows the face video completion results in a practical scenario, where user's face is partially occluded by HoloLens.

5 Conclusion, Limitation and Future Work

We formulate the HMD occlusion removal task in VR communications as a face image completion problem, and propose a novel light-weighted solution that does not rely on hardwares. Given an occluded input image and an occlusion-free reference image, our method automatically synthesizes facial content for the occluded region using the proposed head pose features and face completion network. Results from the network are visually better than those from existing face completion methods. We believe the proposed method is complementary to current hardware-based solutions and provide a novel perspective for HMD removal in VR applications.

Our method, in current form, may not perform well for extreme viewing angles (see Figure 10). This is because the state-of-the-art face alignment method may not able to faithfully estimate landmarks for the large occluded region in such cases. While face landmark detection is not claimed as our contribution, a more robust method is expected. Moreover, for video applications, as our method separately handles each frame without considering temporal coherency, jittering and flicking artifacts can be observed. One possible improvement solution is to design a 3D deep convolutional network that preserves temporal stability. Last, our method does not explicitly manipulate the gaze direction of the result. How to design an end-to-end network that imposes precise controls on gaze is worth investigating.

Acknowledgments

This work was supported by the Research Program of State Key Laboratory of Virtual Reality Technology and Systems (No. VRLAB2019P16) and the Young Talent Program of Beihang University (No. YWF-19-BJ-J-361).

References

[1] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017.

[2] C. Cao, D. Bradley, K. Zhou, and T. Beeler. Real-time high-fidelity facial performance capture. *ACM Trans. Graph.*, 34(4):46:1–46:9, 2015.

[3] S.-Y. Chen, L. Gao, Y.-K. Lai, P. L. Rosin, and S. Xia. Real-time 3d face reconstruction and gaze tracking for virtual reality. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 525–526. IEEE, 2018.

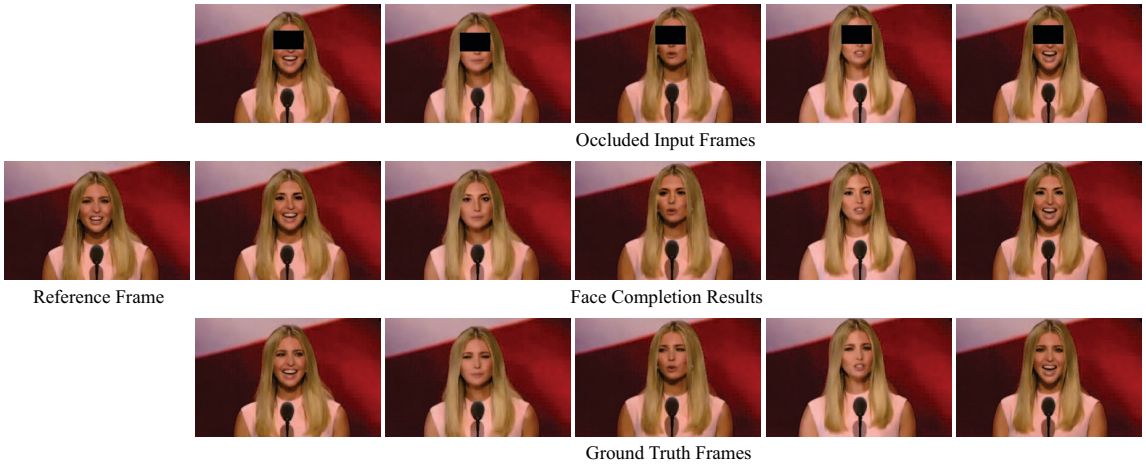

Figure 8: Representative results on video frames. From top to bottom: occluded input frames, our results, and the ground truth frames. The reference image is shown left-most.

Figure 9: Face completion of real video data. The first and second rows are the input frames and our results. The leftmost is the occlusion-free reference image.

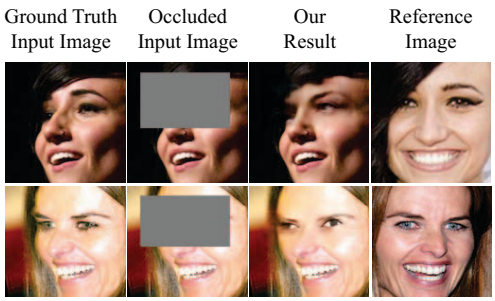

Figure 10: Failure cases. For extreme viewing angles, our method may generate unnatural results that do not accurately reveal the identity.

[4] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[5] B. Dolhansky and C. Canton Ferrer. Eye in-painting with exemplar generative adversarial networks. In *CVPR*, pages 7902–7911, 2018.

[6] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversial nets, 2014. In *Conference on Neural Information Processing Systems*, 2672–2680.

[8] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, pages 87–102, 2016.

[9] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36(4):107:1–107:14, 2017.

[10] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[11] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep video portraits. *ACM Trans. Graph.*, 37(4):163:1–163:14, 2018.

[12] I. Korshunova, W. Shi, J. Dambre, and L. Theis. Fast face-swap using convolutional neural networks. In *ICCV*, pages 3677–3685, 2017.

[13] Y. Li, S. Liu, J. Yang, and M.-H. Yang. Generative face completion. In *CVPR*, pages 3911–3919, 2017.

[14] R. Natsume, T. Yatagawa, and S. Morishima. Fsnet: An identity-aware generative model for image-based face swapping. In *ACCV*, 2018.

[15] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros. Context encoders: Feature learning by inpainting. 2016.

[16] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Trans. Graph.*, 22(3):313–318, 2003.

[17] T. Portenier, Q. Hu, A. Szabo, S. A. Bigdeli, P. Favaro, and M. Zwicker. Faceshop: Deep sketch-based face image editing. *ACM Transactions on Graphics (TOG)*, 37(4):99, 2018.

[18] L. Song, J. Cao, L. Song, Y. Hu, and R. He. Geometry-aware face completion and editing. *arXiv preprint arXiv:1809.02967*, 2018.

[19] M. Takemura and Y. Ohta. Diminishing head-mounted display for shared mixed reality. In *Proceedings of the 1st International Symposium on Mixed and Augmented Reality*, ISMAR '02, pages 149–, 2002.

[20] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Facevr: Real-time gaze-aware facial reenactment in virtual reality. *ACM Trans. Graph.*, 37(2):25:1–25:15, 2018.

[21] J. Thies, M. Zollhöfer, C. Theobalt, M. Stamminger, and M. Niessner. Headon: Real-time reenactment of human portrait videos. *ACM Trans. Graph.*, 37(4):164:1–164:13, 2018.

[22] M. Wang, G.-Y. Yang, R. Li, R.-Z. Liang, S.-H. Zhang, P. M. Hall, and S.-M. Hu. Example-guided style-consistent image synthesis from semantic labeling. In *CVPR*, 2019.

[23] M. D. Zeiler. Adadelta: An adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.

[24] S. Zhang, R. Liang, and M. Wang. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media*, 5(1):105–115, 2019.

[25] Y. Zhao, W. Chen, J. Xing, X. Li, Z. Bessinger, F. Liu, W. Zuo, and R. Yang. Identity preserving face completion for large ocular region occlusion. *arXiv preprint arXiv:1807.08772*, 2018.