Identity Preserving Face Completion for Large Ocular Region Occlusion

Yajie Zhao[1]
yajie.zhao@uky.edu
Weikai Chen[2]
wechen@ict.usc.edu
Jun Xing[2]
junxnui@gmail.com
Xiaoming Li[3]
hit.xmshr@gmail.com
Zach Bessinger[1]
zach.bessinger@gmail.com
Fuchang Liu[4]
20140022@hznu.edu.cn
Wangmeng Zuo[3]
cswmzuo@gmail.com
Ruigang Yang[1]
ryang@cs.uky.edu

[1] Computer Science Department
University of Kentucky
Lexington, KY, USA

[2] Institute for Creative Technologies
University of Southern California
Playa Vista, California, USA

[3] School of Computer Science and Technology
Harbin Institute of Technology
Harbin, China

[4] Hangzhou Institute of Service Engineering
Hangzhou Normal University
Hangzhou, China

Abstract

We present a novel deep learning approach to synthesize complete face images in the presence of large ocular region occlusions. This is motivated by recent surge of VR/AR displays that hinder face-to-face communications. Different from the state-of-the-art face inpainting methods that have no control over the synthesized content and can only handle frontal face pose, our approach can faithfully recover the missing content under various head poses while preserving the identity. At the core of our method is a novel generative network with dedicated constraints to regularize the synthesis process. To preserve the identity, our network takes an arbitrary occlusion-free image of the target identity to infer the missing content, and its high-level CNN features as an identity prior to regularize the searching space of generator. Since the input reference image may have a different pose, a pose map and a novel pose discriminator are further adopted to supervise the learning of implicit pose transformations. Our method is capable of generating coherent facial inpainting with consistent identity over videos with large variations of head motions. Experiments on both synthesized and real data demonstrate that our method greatly outperforms the state-of-the-art methods in terms of both synthesis quality and robustness.

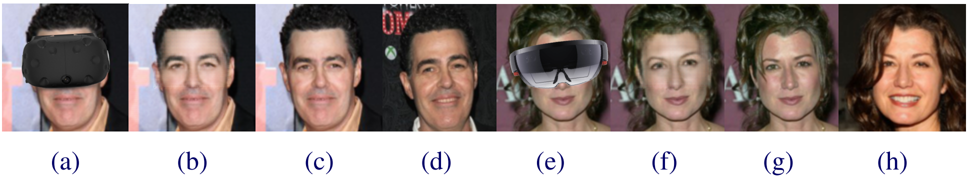

Figure 1: Given a significantly occluded facial image (a) (e), we synthesize the un-occluded face image using our framework with identity preserved (b)(f). (c)(g)show the ground truth image and (d)(h) show the reference images we use of the same person to provide identity information.

1 Introduction

Wearable VR/AR devices provide users the ability to travel freely through physical environments mixed with immersive virtual content, enabling new applications in entertainment, education and telepresence. However, the large occlusion introduced by head-mounted display (HMD) is a huge hindrance for face-to-face communications. Such limitation could prevent the adaptation of VR/AR technologies in areas, such as teleconferencing, in which eye contact and facial expressions are crucial elements in effective team communication and negotiation tactics.

To enable better face-to-face like communications when wearing a HMD, researchers have developed techniques to capture a wearer's expressions to drive a digital avatar [14, 19, 23]. Although some impressive results have been demonstrated, the visual representation is only used for a "talking head" in the VR setting and is limited in quality and details. Instead of driving virtual avatars, another research direction is to inpaint the occluded regions with plausible faces. Nevertheless, inferring the occluded content introduced by VR goggles is particularly challenging as over half of the face is obstructed in the most cases. [3, 33] tried to synthesize the missing texture using personalized database, but requires dedicated capturing setup that makes their methods hard to generalize.

Although some of the state-of-the-art algorithms [16] and [12] are able to produce plausible face image with large occlusions. Their inputs are required to be frontal, aligned and the identity is usually not preserved during the completion. Such limitations make them infeasible in applications like headset removal, as identity is required to be preserved while the face pose is likely to be changing.

We thus present a novel deep learning approach that can not only fill in the large occluded regions with plausible contents but also provide control over the restored face identity and face poses as shown in Figure 1. The user could specify the desired face identity by providing an arbitrary occlusion-free face image of the target subject as reference. In addition, by inputting a pose map, our approach could generate facial structures consistent to the intended face orientation. These advances are enabled by a generative network that is optimized with dedicated constraints to regularize the synthesis process. To inpaint the occluded region with facial content that is visually similar to the input reference image, we introduce a novel reference network that imposes an identity prior onto the searching space of generator. The identity prior is extracted from the referenced identity and penalizes stylistic deviation between the generated result and the input reference image. At most cases, the reference image is prone to have different pose, illumination and background with the input. To obtain a spatially-coherent result, we regularize the generator using two discriminators: a global discriminator that enforces context consistency between filled pixels with surrounding background, and a pose discriminator that regularizes the high-level postural errors. The pose map serves as both the input of generator and the condition of pose discriminator. By observing the ground-truth face pose, the pose discriminator penalizes unreal pose transformations produced from the generator.

Compared with the previous state-of-the-art methods, our approach is more advantageous in the following aspects. 1) Our method provides significantly better results in the presence of large occluded regions, *e.g*. obstruction from large HMDs. 2) We propose the first face inpainting framework that could explicit control the recovered face identity, which makes identity preserving possible in headset removal. 3) Our approach also offers the editing of face poses in the restored content. To the best of our knowledge, this is the first work that could achieve realistic pose-varying face completion in videos.

2 Related Work

Synthesizing the missing portion of a face image could be formulated as an inpainting problem, which is first introduced in [2]. To obtain a faithful reconstruction, content prior is usually required, which comes from either other part of the same image or an external dataset. The former method generates reasonable inpaintings under specific assumptions, such as repetitiveness of texture [5], spatial smoothness in the missing region [22] or planar objects [10]. However, these methods are prone to fail when completing images with structured content. The data-driven methods leverage learnt features from database to infer the missing content [6, 8, 9, 18, 21, 25, 26, 29]. In particular, the authors in [9, 18, 25] generate complete image automatically by using a feature dictionary.

Deep neural network based methods [6, 8, 26] hallucinate the missing portion of the images by learning through the background texture. However, the early attempts tend to generate blurry results and have no control over the semantic meaning of generated result. More recently, several GAN frameworks have been proposed to address this issue [12, 13, 17, 21, 28, 29, 30]. GANs have been shown to perform well in generating realistic appearing images. [12, 17] solve the general face completion problem by training a model with global and local discriminators. These discriminators ensure that the generated face appear realistic. In the face inpainting work of Yeh *et al.* [30], they search the closet encoding in the latent image manifold to get an inference of how the missing content should be structured, which predicts information in large missing regions and achieve appealing results. None of the existing face inpainting approaches is capable of preserving the identity, which makes it infeasible to be applied in the headset removal applications.

The identity-preserving problem has been explored in related tasks [11, 15, 24, 31], e.g. attributes transfer, frontalization and face recognition. In particular, pose code has also been introduced in [24, 31] to resolve the identity ambiguity. However, trivially applying the above-mentioned approaches would fail in our case as none of these works has considered large occlusion, *e.g*. entirely blocked upper face, in their formulation. We show that by jointly learning features from an arbitrary image of the target identity and a control pose map can significantly improve the inpainting performance while achieving additional control of face identity and head pose, which, for the first time, enables inpainting over a dynamic sequence with large head pose variations.

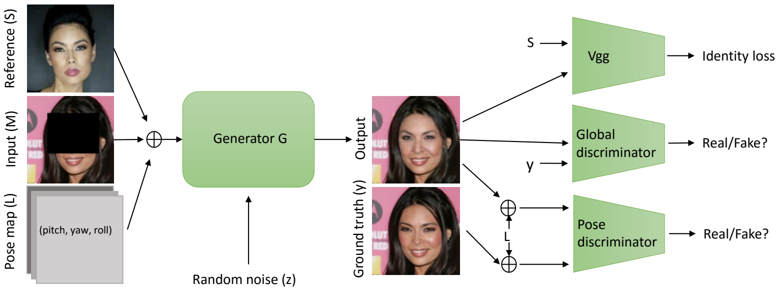

Figure 2: Our network architecture

3 Identity Preserving Face Completion

To faithfully inpaint the missing region with realistic content while resembling the input identity under a pose constraint, we propose an architecture that consists of a generator and two discriminators as illustrated in Figure 2.

3.1 Generator

To inpaint large occluded regions with controllable face identity and head pose, the generator G of our network takes three inputs: the occluded face image M, an occlusion-free image of a reference identity S and a pose map L that controls the head pose of generated result. The pose map L is a constant-value color image with its three channels encoded by normalized pitch, yaw and roll angles that define the intended face orientation. Starting from a random variable z, we progressively optimize the generator so that it could learn the mapping from a normal distribution to an image manifold $\bar{z}$ that is close to both groundtruth y and the reference image S under the pose constraint L. We formulate the process of finding a recovered encoding $\bar{z}$ as a conditioned optimization problem. In particular, $\bar{z}$ is optimized via solving the following equation:

$$\bar{z} = \arg\min_{z} \{\mathcal{L}_r(z|M,S,L) + \mathcal{L}_{\text{id}}(z,S)\} \tag{1}$$

where $\mathcal{L}_r$ indicates the reconstruction loss and $\mathcal{L}_{\text{id}}$ denotes an identity loss that penalizes the deviation from the referenced identity. As L_2 loss empirically leads to blurry output and L_1 loss performs better on preserving of high-frequency details, we use the L_1 loss for measuring the reconstruction error between the generated result and the groundtruth image:

$$\mathcal{L}_r = \|y - G(z|M,S,L)\|_1 \tag{2}$$

Though conditioned on the reference image, only reconstruction loss is not sufficient for ensuring visual similarity with the referenced identity. To achieve identity preservation, we propose to add an identity loss by introducing a *reference network* R that extracts high-level features from the generated result and reference image. We utilize the pre-trained VGG Face network [20] as our feature extractor. In particular, we use the $FC6$ feature for both input images. We define the identity loss as the L_2 distance between the extracted feature vectors:

$$\mathcal{L}_{\text{id}} = \|f(G(z|M,S,L)) - f(S)\|_2 \tag{3}$$

where f represents the non-linear feature extracting function learnt by R.

Note that we do not require the referenced identity image to be pose-aligned with the groundtruth. But our model can still accurately capture the semantic features from the reference image. As demonstrated in Figure 4, the identity-dependent features have been successfully transferred to the generated image. We thus interpret the reference network as a regularizer that imposes an identity/style prior on the manifold of generated images. The proposed network can not only improve the synthesis quality but also stabilize the output to enhance the temporal coherence when dealing with dynamic sequences, e.g. videos.

3.2 Discriminator

Though the generator can synthesize the missing content with low reconstruction and identity errors, there is no guarantee that the generated image is realistic and consistent with surrounding background. Discriminator serves as a binary classifier that distinguishes real and fake images so that it helps improve the synthesis quality. To encourage photorealism and effective control of face pose, we introduce two discriminators to supervise the generator.

We first introduce a global discriminator D to justify the fidelity and coherence of the entire image. The rationale for introducing a global discriminator is that the inpainted content should not only be realistic but also spatially coherent with surrounding context. In addition, the global discriminator should impose constraints on forming semantic valid facial structures. In particular, we formulate the global discriminator loss function as below:

$$\mathcal{L}_{\text{global}} = \min_{G} \max_{D_g} \; \mathbb{E}_{x \sim p_{\text{data}}(x)} \left[\log D_g(x, M) \right] + \mathbb{E}_{z \sim p_z(z)} \left[\log\left(1 - D_g(G(z|M,S,L), x)\right) \right]. \tag{4}$$

where $p_{\text{data}}(x)$ and $p_z(z)$ represent the distributions of real data x and noise variables z respectively. The global discriminator is sufficient for synthesizing occluded faces with fixed face pose. However, in our application scenario, where the HMD wearer is likely to rotate his/her head while talking, our network should be robust to variable face poses. That means the generated content should have facial structures oriented consistently with the input head pose. We therefore propose an additional pose discriminator D_{pose} to distinguish the faithfulness of synthesized result given the pose constraint. In particular, the pose loss is defined as follows:

$$\mathcal{L}_{\text{pose}} = \min_{G} \max_{D_p} \; \mathbb{E}_{x \sim p_{\text{data}}(x)} \left[\log D_p(x, L) \right] + \mathbb{E}_{z \sim p_z(z)} \left[\log\left(1 - D_p(G(z|M,S,L), L)\right) \right]. \tag{5}$$

We condition the loss of pose discriminator on the pose map L so that the input pose map would have more accurate control over the inpainted result. However, unlike the global discriminator that back-propagates the gradient over the entire image, the pose discriminator only supervises the loss gradients for the missing region.

Therefore, the overall loss function of our network is defined by:

$$\mathcal{L} = \lambda * \mathcal{L}_r + \mu * \mathcal{L}_{\text{id}} + \alpha * \mathcal{L}_{\text{global}} + \gamma * \mathcal{L}_{\text{pose}} \tag{6}$$

where λ, μ, α, and γ are the weights for the reconstruction loss, identity loss, global discriminator and pose discriminator loss, respectively.

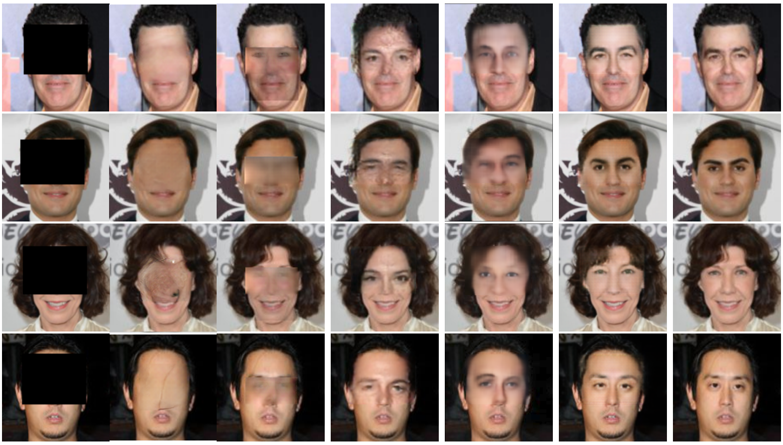

Figure 3: Comparison with state-of-the-art inpainting frameworks.

3.3 Architecture

In our experiment, we adopt U-Net architecture with skipped connections as our generator. Specifically, we concatenate the i-th layer onto the $(N-i)$-th layer, where N is the total number of layers, to avoid information loss caused by the bottleneck layer. On the discriminating side, we use PatchGAN for both global and pose discriminator.

4 Implementation Details

We use images from MS-celeb-1M [7] to construct our training data. Our model is trained using 476 identities and 8000 pair of images (the occluded face image and its reference identity image). To prepare the data, MTCNN [32] is applied to detect the landmarks and bounding boxes. We then scale all the images of the dataset to $128*128$ and align them by registering the nose tip. The loss functions are optimized using the Adam optimizer, with a learning rate of 0.0002 and $\beta_1 = 0.5$. We train the network for 100 epochs. Our framework is implemented using Torch [4]. In all experiments, we set our loss hyperparameters as $\lambda = 1$, $\mu = 100, \alpha = 100$, and $\gamma = 70$. The momentum is set to 0.9 in our training process.

Runtime. We implement our model with Torch [4] on a platform of Intel E3 CPU, 3.30GHz and Nvidia GTX-1080 GPU. We can reconstruct face images with size 128×128 at a frame rate of 20 Fps.

5 Experimental Results

5.1 Face Identity Control

In this section, we evaluate the effectiveness of the proposed face identity control. Figure 4 demonstrates the cross identity experiments, where the image of another identity is fed into the network as reference image. As seen in Figure 4, the referenced identity differs significantly from the original identity in terms of appearance and even genders. However, our model can still generate high-fidelity result with spatial coherence while capturing the high-level identity-dependent features, *e.g*. thick eyebrows, eye color, of the referenced identity.

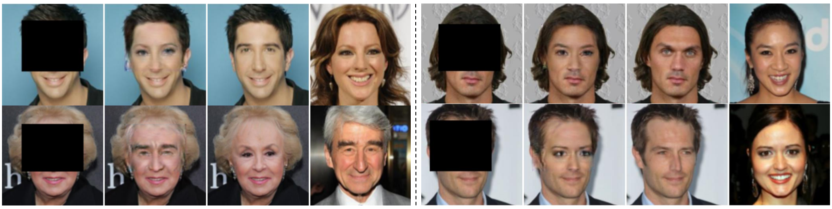

Figure 4: Cross identity experiments. From left to right: inputs, generated image, ground truth, reference images of another identity.

Table 1: Comparison of Face Verification

	ours (%)	Li (%)	Iizuka (%)
Compare with groundtruth	93.2	57.3	45.9
Compare with reference	87	42.8	34.7

Quantitative Analysis To further quantify the performance of identity preservation, for each synthesized result, we apply the OpenFace [1] to verify the identity similarity of our result compared to the ground-truth and reference image used in our network. In particular, the OpenFace will generate a binary output (0 or 1) for indicating if its two input images capture the same identity. We compare the performance of Li *et al.* [17], Iizuka *et al.* [12] and our algorithm using 588 images pairs(the input and reference image are the same person). As demonstrated in Table1, our method significantly outperforms Li *et al.* [17] and achieves a 93% pass rate when comparing with the groundtruth, indicating the efficacy of our method in preserving the target identity.

5.2 Head Pose Editing

In this section, we validate the effectiveness of the proposed pose editing component. Figure 5 shows how the pose variation influences the reconstruction of the face images. By providing different pose inputs, our reconstructed facial structures will be aligned accordingly. The first row of Figure 5 shows the impact of pitch variance on reconstructed results. By gradually increasing the pitch value, the synthesized eye will move up accordingly. The similar control effect is manifested in tuning the yaw values as shown in the second row of Figure 5. As the face regions outside the mask remain fixed, it is more natural that only the correct pose input will lead to the most coherent and clear results. As seen from our result, only the closer poses (highlighted in red) generates visually better results, suggesting the accuracy of our pose control.

The introduction of pose editing component ensures our model to perform face inpainting in dynamic sequence with time-space coherence. In Figure 6, we show the reconstructed frames with different head poses from a video sequence by using only one frontal reference. Regardless of the large variations of poses, our network can stably reconstruct appealing results. We provide video results at https://youtu.be/4qGaARE8ob4 to better evaluate our method.

5.3 Ablative Analysis

To access the efficacy of each introduced loss, we experiment on three combinations of losses: L_1+GAN, L_1+GAN+ID and L_1+GAN+ID+Pose. Figure 7 shows the comparison of the above networks. In general, L_1+GAN tends to generate blurry results and fails to capture

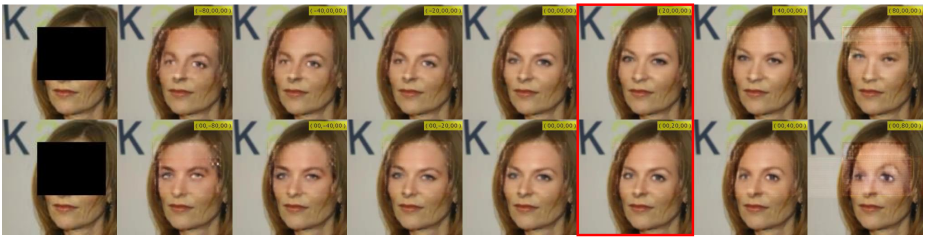

Figure 5: Impacts of pose variation on the face reconstruction. The ground truth pose of the input image (first column) is $(pitch = 15, yaw = 20, roll = 0)$, and first/second rows show the effects of different pitch/yaw values on the reconstruction results. As we can see, the close poses (in red box) tend to generate visually better results.

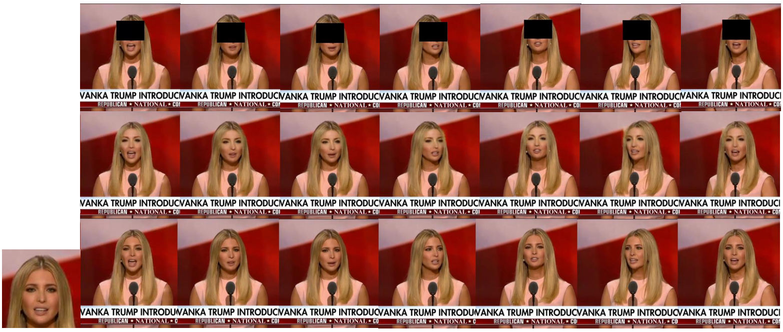

Figure 6: Video example. From top to bottom: input video frames, generated outputs from our network, and the ground truth. The reference image is provided left-most.

spatial coherency.L_1+GAN+ID demonstrates better performance on preserving the identity, although the result is still blurry and mis-aligned in pose. L_1+GAN+ID+Pose, which is our proposed method, is capable to generate images with sharper details while faithfully capturing the referenced identity. The results indicate that the pose control component acts as an implicit alignment prior to register different features to reduce the blurness for each semantic part. Table 2 shows the quantitative evaluation on the test set. Our proposed network outperforms the other methods in both PSNR and SSIM.

6 Comparisons

Comparisons with other methods We compare our result with other state-of-the-art inpainting frameworks. As shown in Figure 3, Pathak *et al.* [21] smoothly fill the missing part without any semantic meaning. Though Yeh et al. [30] can produce content with semantic meanings, their results tend to be blurry and fail to be spatially coherent with surrounding context. Li *et al.* [17] and Iizuka *et al.* [12] demonstrate sharper results but the results are still blurry and the generated facial features tends to be distorted and appear unnaturally with unmasked regions. Comparing to other methods, our approach is capable of generating high-fidelity facial details with coherent blending with backgrounds. In addition, we successfully

Table 2: Quantitative Evaluation on different Networks

	Yeh [30]	Li [17]	Iizuka [12]	L_1+ GAN + ID	Ours
PSNR	18.87	21.69	23.14	23.81	24.9
SSIM	0.79	0.78	0.82	0.83	0.87

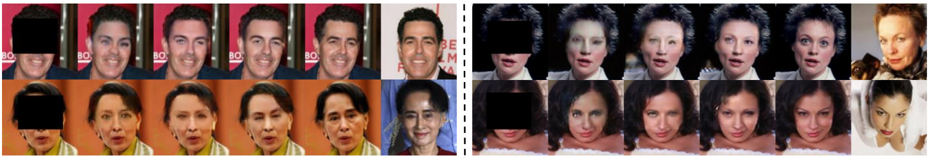

Figure 7: The ablative analysis. From left to right: input, L_1+GAN, L_1+GAN+ID, L_1+GAN+ID+pose, ground truth, and reference image.

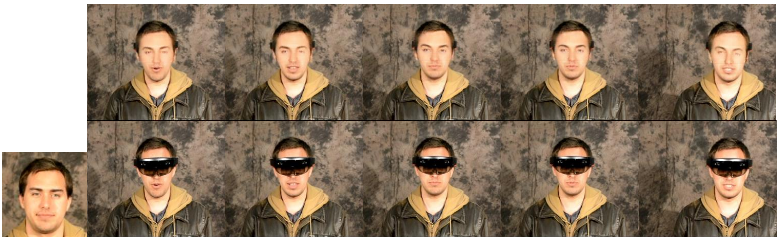

Figure 8: Reconstruction of real video data with frontal head pose. The first and second rolls are the selected frames reconstructed by our network and the ground truth. The leftmost image is used as reference for all the frames.

preserve the identity, providing result close to the ground truth.

Test on images with real occlusion. We test our network on video sequences in which the subjects are wearing HMDs. As we assume known head poses for our network. We need first to extract the head poses from occluded images. However, many HMDs, like HTC vive, provide real-time tracking of the head pose, which could be converted to our pose input via a simple calibration step. We also train a pose prediction network using the synthetic data with known pose information. The network consists of 5 convolutional layers to extract the high-level features from the input image, and two fully connected layers to regress the feature into pose. In particular, our convolution part is same as the content prediction network of [27], which is trained to inpaint the missing content. In Figure 8, we show a video result with nearly frontal view, where we assume that pose is fixed to $(0,0,0)$. As seen from the results, our network produces stable results when the pose changes slightly. In Figure 9, we test our network on a video in which the subject is wearing a HTC vive VR headset and talking with large variations of head poses. Despite the large head movement, our network can still generate very promising results. Although our results of real data contain artifacts, the improvement is significant compared to Li *et al.* [17] and Iizuka *et al.* [12].

7 Conclusion, Limitation and Future work

We present a novel learning-based approach for face inpainting with favorable property of preserving the identity of a given reference image. Furthermore, our approach offers flexible pose control on the reconstruction results, making it possible to faithfully restore facial details in occluded video sequences with large face pose variations. These two properties provide insight into solving the headset removal problem, which attracts increasing attention due to the surge of VR/AR techniques. Our network, in current form, cannot handle well extreme viewing angles and expressions (failure cases can be found in supplementary materials). However, we believe that by including such cases in training dataset, the robustness of our network can be further improved. In the video inpainting results, jittering can be

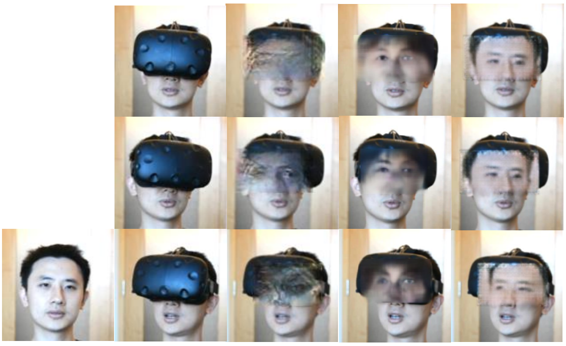

Figure 9: Reconstruction of real video data with large head pose variation when wearing VR/AR headsets. From the left to the right columns are: *One single reference image*, *Inputs*, *Li* et al. *[17]*, *Iizuka* et al. *[12]* and *Ours*. The single reference image is used for all frames.

observed in the transition between different frames as temporal coherency is not explicitly constrained in our formulation. Its worth investigating in the future work to add such additional constraint in our current framework. The results of real data generated by our network still have some artifacts around the mask boundary. This is due to the fact that the lower face usually has shadows cast by the HMDs. One possible solution is to synthesize shadows for the training data. It would also be an interesting future work to incorporate a pose estimation network to enable an end-to-end face inpainting network for videos.

8 Acknowledgements

We gratefully acknowledge the support of NSFC grant(61502133). We would like to show our gratitude to Nathan Jacobs, Xinyu Huang for sharing their pearls of wisdom with us during the course of this research, and we thank Qingguo Xu for assistance with experiments and comments that greatly improved the manuscript.

References

[1] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[2] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424. ACM Press/Addison-Wesley Publishing Co., 2000.

[3] Xavier P. Burgos-Artizzu, Julien Fleureau, Olivier Dumas, Thierry Tapie, FranÃğois Le Clerc, and Nicolas Mollet. Real-time expression-sensitive hmd face reconstruction. In *SIGGRAPH Asia Technical Briefs*, 2015.

[4] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, number EPFL-CONF-192376, 2011.

[5] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1033–1038. IEEE, 1999.

[6] Alhussein Fawzi, Horst Samulowitz, Deepak Turaga, and Pascal Frossard. Image inpainting through neural networks hallucinations. In *Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), 2016 IEEE 12th*, pages 1–5. IEEE, 2016.

[7] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer, 2016.

[8] Kushagr Gupta, Suleman Kazi, and Terry Kong. Deeppaint: A tool for image inpainting.

[9] James Hays and Alexei A Efros. Scene completion using millions of photographs. In *ACM Transactions on Graphics (TOG)*, volume 26, page 4. ACM, 2007.

[10] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Image completion using planar structure guidance. *ACM Transactions on Graphics (TOG)*, 33(4): 129, 2014.

[11] Rui Huang, Shu Zhang, Tianyu Li, Ran He, et al. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. *arXiv preprint arXiv:1704.04086*, 2017.

[12] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 36(4):107, 2017.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

[14] Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. Facial performance sensing head-mounted display. *ACM Transactions on Graphics (TOG)*, 34(4):47, 2015.

[15] Mu Li, Wangmeng Zuo, and David Zhang. Deep identity-aware transfer of facial attributes. *arXiv preprint arXiv:1610.05586*, 2016.

[16] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[17] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[18] Julien Mairal, Michael Elad, and Guillermo Sapiro. Sparse representation for color image restoration. *IEEE Transactions on image processing*, 17(1):53–69, 2008.

[19] Kyle Olszewski, Joseph J. Lim, Shunsuke Saito, and Hao Li. High-fidelity facial and speech animation for vr hmds. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia 2016)*, 35(6), December 2016.

[20] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[21] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. *arXiv preprint arXiv:1604.07379*, 2016.

[22] Jianhong Shen and Tony F Chan. Mathematical models for local nontexture inpaintings. *SIAM Journal on Applied Mathematics*, 62(3):1019–1043, 2002.

[23] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Facevr: Real-time facial reenactment and eye gaze control in virtual reality. *arXiv preprint arXiv:1610.03151*, 2016.

[24] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, volume 3, page 7, 2017.

[25] Oliver Whyte, Josef Sivic, and Andrew Zisserman. Get out of my picture! internet-based inpainting. In *BMVC*, pages 1–11, 2009.

[26] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In *Advances in Neural Information Processing Systems*, pages 341–349, 2012.

[27] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. *CoRR*, abs/1611.09969, 2016. URL http://arxiv.org/abs/1611.09969.

[28] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. *arXiv preprint arXiv:1611.09969*, 2016.

[29] Raymond Yeh, Chen Chen, Teck Yian Lim, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with perceptual and contextual losses. *arXiv preprint arXiv:1607.07539*, 2016.

[30] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5485–5493, 2017.

[31] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. *arXiv preprint arXiv:1704.06244*, 2017.

[32] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10): 1499–1503, Oct 2016. ISSN 1070-9908. doi: 10.1109/LSP.2016.2603342.

[33] Yajie Zhao, Qingguo Xu, Xinyu Huang, and Ruigang Yang. Mask-off: Synthesizing face images in the presence of head-mounted displays. *arXiv preprint arXiv:1610.08481*, 2016.