Attention based Occlusion Removal for Hybrid Telepresence Systems

Surabhi Gupta
IIIT Hyderabad, India
surabhi.gupta@research.iiit.ac.in

Ashwath Shetty
IIIT Hyderabad, India
ashwath.shetty@research.iiit.ac.in

Avinash Sharma
IIIT Hyderabad, India
asharma@iiit.ac.in

Figure 1. Our proposed approach can reconstruct high-quality unoccluded image from a given occluded face image.

Abstract*—Traditionally, video conferencing is a widely adopted solution for remote communication, but a lack of immersiveness comes inherently due to the 2D nature of facial representation. The integration of Virtual Reality (VR) in a communication/telepresence system through Head Mounted Displays (HMDs) promises to provide users with a much better immersive experience. However, HMDs cause hindrance by blocking the facial appearance and expressions of the user. We propose a novel attention-enabled encoder-decoder architecture for HMD de-occlusion to overcome these issues. We also propose to train our person-specific model using short videos of the user, captured in varying appearances, and demonstrated generalization to unseen poses and appearances of the user. We report superior qualitative and quantitative results over state-of-the-art methods. We also present applications of this approach to hybrid video teleconferencing using existing animation and 3D face reconstruction pipelines. Dataset is available at this *website.

***Keywords*-face image inpainting, facial de-occlusion, HMD removal, virtual reality**

I. Introduction

Globalization has led to an acute need for tele-interactions for effective communication that has been further boosted due to the current pandemic situation across the world. Traditionally, video conferencing is a widely adopted solution for telecommunication, but it lacks realism due to the 2D nature of facial representation. Virtual Reality (VR) based telepresence system provides a better immersive experience for remote conversation and collaboration. Nevertheless, HMDs significantly occlude the user's face, hindering facial appearance capture, including gaze and expressions. Therefore, HMD removal in images is vital for improving the user experience.

Traditionally, Analysis-by-Synthesis techniques for HMD de-occlusion proposed in the literature [17] animate parametric face models such as 3DMMs [1] using features extracted from an HMD occluded input image. However, such models often generate overly smooth geometrical details and compromise the realism of facial appearance. On the contrary, recent facial avatar-based models [11] achieve photorealistic results for HMD de-occlusion. However, avatar modeling methods require a large amount of calibrated multi-view data of a single user in different poses and expressions for avatar creation. [11] uses a setup consisting of 40 machine vision cameras capable of synchronously capturing 5120 $\times$ 3840 images at 30 frame per second (FPS). Thus, such avatar creation is non-trivial and challenging for a large user base to scale up. Additionally, it is a one-time process for each user. Therefore, such parametric model or avatar-based techniques have a significant limitation: they lack the user's actual appearance during the interaction (i.e., unable to model the everyday appearance of the user), hindering

2022 19th Conference on Robots and Vision (CRV) | 978-1-6654-9774-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/CRV55824.2022.00029

978-1-6654-9774-9/22/$31.00 ©2022 IEEE
DOI 10.1109/CRV55824.2022.00029

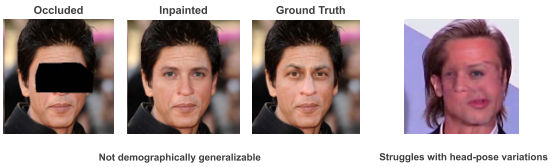

Figure 2. Failure cases of LaFIn [23]

user experience.

HMD de-occlusion problem can also be posed as a face completion/inpainting problem. Existing face completion methods in literature [18], [23] attempt to learn a single inpainting network over a large training population, hoping for good generalization on unseen face images. However, these methods frequently suffer from issues like loss of identity and fail to generalize with even minor variations in head pose, as shown in Figure 2. Another set of methods in [13], [18] uses a reference image along with the occluded image to fill a masked region and preserve identity. However, as shown in Figure 2, their work fails to generalize with non-frontal head-poses. [13] requires additional information (depth and mask) and does not generate the entire face (with hairs, ears, and background), which hinders user experience.

The primary research challenge with the face completion/inpainting task comes from its ill-posed nature as a significant part of the face is occluded by HMD. Learning a single common face de-occlusion network with the capability to hallucinate diverse expressions in varying appearances and head poses across a large set of human faces is difficult to achieve. It is due to the broad space of facial geometry and appearance as well as the highly subjective way of articulating expressions/emotions across individuals [2]. Additionally, in the context of VR teleconferencing applications, the desired solution should be scalable, requiring minimal efforts and hardware at the user's end. An additional desired characteristic might be regarding integration ability in a hybrid VR teleconferencing setup where users with only video capability should also participate as in regular video conferencing.

To overcome these challenges, we propose to tackle this problem in a person-specific setting where we aim to train a dedicated model for each user to learn user-specific appearance, head-pose, gaze, and facial expression traits. The input to our method is a video frame with HMD occluded face, and our model generates a de-occluded plausible face by removing the occlusion. To achieve this, we introduce novel attention enabled encoder-decoder architecture and a novel training strategy to train our person-specific model using short videos (1-2 minutes) of the user. The video captures the varying appearance of the user with a variety of head poses and facial expressions without HMD occlusion. As a part of our training strategy, we first train the encoder-decoder module (sans attention) on large face image datasets to learn generic face appearance features. Subsequently, we finetune it on unoccluded user videos. Finally, we finetune our full model (encoder-decoder with attention) on the same training data with synthetic HMD masks. It allows our model to learn the person-specific facial geometry and expression traits and help generate occluded areas with varying appearances, poses, and expressions.

Our attention module allows the network to preserve the high-frequency appearance and background details (like hairs, wall texture, etc.) from the unoccluded part of the input HMD occluded image while generating the plausible facial appearance for the occluded part. Our novel mask-loss function helps the model to emphasize the occluded region. Figure 1 shows high-quality de-occlusion achieved by our method. It is important to note that learning a person-specific model is not equivalent to overfitting on a specific user as there is a significant change in user appearance, background, and lighting across sessions. Similar person-specific model learning has been successfully explored in this context [11] as well as another related context of egocentric frontal face recovery [5]. We conduct thorough empirical evaluation and report superior qualitative and quantitative results of our proposed method w.r.t. state-of-the-art methods. In addition to this, we also demonstrate the broader applicability of our proposed HMD de-occlusion method with two real-world use cases. First, we can use it to build hybrid VR systems by integrating it with video-driven face animation solutions such as [16], [19]. We also show how to integrate our method with a 3D face reconstruction pipeline to generate 3D face video for the VR teleconferencing system as a second use case. To summarise, we make the following contributions:

1) We present a deep learning framework for person-specific HMD de-occlusion. Our method does not rely on high-end hardware (e.g., HMD with gaze tracking) and calibrated data from the user (for avatar creation).
2) We introduced a novel attention module knitted with the encoder-decoder architecture that can use background and appearance details from the input image and allows the model to focus on inpainting the occluded region with plausible details.
3) We collected a small dataset of multiple users in different appearances, facial expressions, and head-poses that we intend to release to the academic community.
4) We present applications of our model where it can be integrated with neural animation models such as [16], [19] to animate an occluded video and can also be used to recover 3D face from occluded input.

II. Related Work

Our approach is an inpainting method that learns to fill in user-specific details. It is related to traditional inpainting methods and recent approaches that use a reference image to fill an occluded region faithfully. In the following section, we discuss the most relevant literature in detail.

A. Person Specific Models

Recent deep-learning advances in vision and graphics have led to the rise of personalized models that are animated/rendered using deep learning. The closest work to ours is [5]. They use a video-to-video GAN [7], which takes in egocentric frames of a person and generates the corresponding frontal view. These methods show the ability of person-specific models to capture high-frequency details. [11] learns auto-encoder network to predict view conditioned texture and mesh geometry from HMD occluded input. [6] trains a dynamic neural radiance field model on a short video (2-3 min) of the person with different expressions and poses. However, they require calibrated multi-view data from the user, adding additional hardware constraints. Animating these models is also expensive. Our approach requires a short uncalibrated video from the user for training and is considerably lightweight compared to avatar-based methods.

B. Image Inpainting Methods

Image inpainting describes the task of filling missing image regions with realistic content. Recent works [20], [21], [23] train a conditional GAN [8] on a face dataset as a solution to this problem. These methods show impressive generalization to examples with frontal head pose and arbitrary occlusion. Another method, EdgeConnect [24] fills the missing region using edges as prior. However, these methods are biased to their training distribution as they do not generalize well to even slightly non-frontal head poses and suffer from a loss of identity. We train our method only on images of the same person with considerable variations in head poses and expressions that overcome the challenges of identity loss and generalization to various head poses.

C. HMD Removal

Exemplar guided image inpainting methods such as [13], [18] propose an image-based approach to HMD de-occlusion. They use a reference image to guide the inpainting procedure and learn a general model for the task. However, [18] fails to work well with cases of significant pose variations between the reference and occluded image. Also, [13] train and evaluate on synthetic data with additional depth information, which may not work or be available in a real-world teleconferencing scenario. We train and evaluate our model on real-world conversations and scenarios and show our model's ability to generalize to unseen appearances.

III. Methodology

A. Overview

The primary focus of our work is to learn a personalized model for face de-occlusion, particularly as an application in VR teleconferencing, where the face is partially occluded due to HMD. To tackle this, we formulate the face de-occlusion problem as an image inpainting task. Given an occluded face image as an input X_{occ}, our network aims to hallucinate the missing region with plausible and perceptually consistent facial details in order to reconstruct the generated unoccluded image, X_{rec} against the ground truth unoccluded image, X_{gt}.

Inspired by the autoencoder architecture proposed in [3], we use an novel attention enabled encoder-decoder framework with generative capabilities that learns to reconstruct high-fidelity unoccluded faces from HMD occluded input images. Additionally, we also propose a novel mask-based loss function and a novel training strategy to learn our model. Figure 3 shows the outline of our proposed architecture.

B. Proposed Architecture

1) Encoder-Decoder Module: Our encoder-decoder module comprises a stack of ResNet and inverted ResNet blocks. Each ResNet block consists of a set of convolutions with residual connections. For the inverted ResNet block, the first convolution in the ResNet block is replaced by a 4×4 deconv layer. We also provide the additional generative capability to the network using an adversarial loss. The encoder learns a 256-dimensional feature representation of the input image. This bottleneck representation is subsequently fed to the decoder network to reconstruct the target image.

2) Attention Module: Inspired from existing literature on attention-based learning strategies as proposed in [15], [10], we append our encoder-decoder module with an attention module. We perform spatial attention by taking the encoder output from the second layer, F_{enc} of spatial dimension 64×64 and perform a channel-wise concatenation with the corresponding decoder layer output F_{dec} of the same spatial dimension. We feed it to our attention module, which subsequently generates attention maps of the same dimension as shown in Figure 3. We then decouple these attention maps and use them for a weighted fusion of respective feature maps (i.e., F_{enc} and F_{dec}). The fused feature maps are fed downstream to convolution layers to reconstruct the de-occluded face image.

Such attention-based spatial feature map fusion allows our network to preserve high-frequency appearance/background details (like hairs, wall texture, etc.) from the visible part of the input image while generating a plausible facial appearance for the occluded part. These attention maps can be learned using fully convolutional networks. As shown in Figure 3, the attention module consists of $Conv(4m, 3)$, $Conv(4m, 3)$, $Conv(8m, 3)$ and $Conv(2m, 3)$, where m denotes the base number of filters and $Conv(m, k)$ denotes a convolutional layer with output number of channels m and kernel size k. The final output with $2 * m$ channels is then split into two, $Attn_{enc}$ and $Attn_{dec}$, each of m channels and spatial dimension of 64×64. This acts as a attention mask for the inputs, F_{rec} and F_{dec} which is then fused again using a channel-wise summation according to Equation 1.

$$F_{fused} = F_{enc} * Attn_{enc} + F_{dec} * Attn_{dec} \quad (1)$$

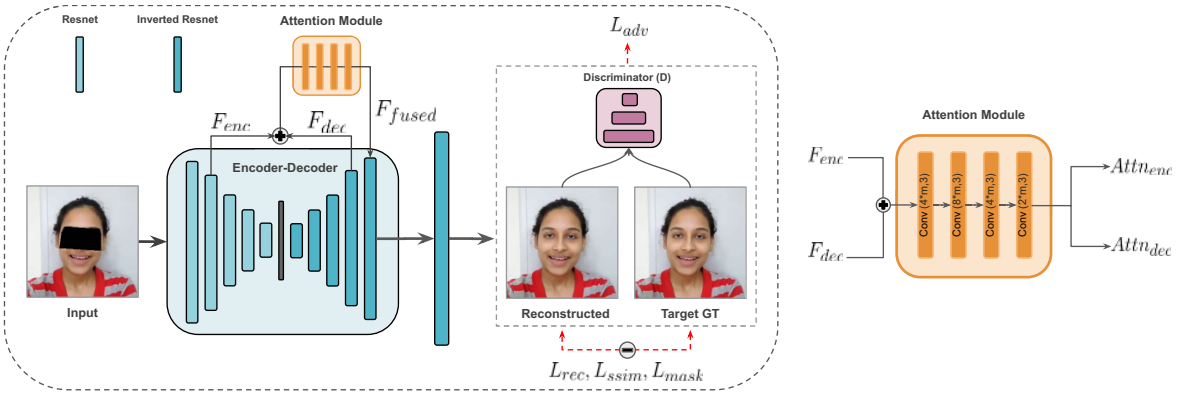

Figure 3. An overview of our proposed facial de-occlusion network.

3) Loss Function: We employ a combination of four different loss functions as our training objective. In order to penalize reconstruction errors, we use pixel-based $L1$ loss.

$$L_{rec} = \|X_{gt} - X_{rec}\|_1 \tag{2}$$

However, using only $L1$ reconstruction loss produces blurry outputs. To overcome this, we add a discriminator, D in the architecture to compute the adversarial loss. This adversarial loss term forces the encoder-decoder to reconstruct high-fidelity outputs by sharpening the blurred images. For D, we adopt the architecture of the DCGAN discriminator [14].

$$L_{adv} = log(D(X_{gt})) + log(1 - D(X_{rec})) \tag{3}$$

We also use SSIM based structural similarity loss, as defined in [3], that helps to improve the alignment of high-frequency image elements to stabilize the adversarial training.

$$L_{ssim} = SSIM(X_{rec}, X_{gt}) \tag{4}$$

To further improve the quality of reconstruction in the HMD occluded area of the generated image, we propose a novel *mask-based loss*. Here, we use the binary mask image as an additional supervision to the network along with input image while training. Minimizing this loss helps the model to emphasize more on quality of reconstruction in the masked region. This also helps to mitigate the blinking artifacts around the eye region for stable reconstructions.

We formulate the mask-based loss function as:

$$L_{mask} = \|I_{mask} \odot I_{gt} - I_{mask} \odot I_{rec}\|_1 \tag{5}$$

where, I_{mask} refers to single channel binary mask image where white pixels (1) correspond to occluded region and black pixels (0) correspond to the remaining unoccluded region and $\odot$ is element-wise multiplication. Thus, the final training objective loss function can be written as,

$$\begin{aligned} L_{final} =& \lambda_{rec} * L_{rec} + \lambda_{adv} * L_{adv} + \lambda_{ssim} * L_{ssim} \\ &+ \lambda_{mask} * L_{mask} \end{aligned} \tag{6}$$

where, $\lambda_{rec}, \lambda_{adv}, \lambda_{ssim}$ and λ_{mask} are the corresponding weight parameters for each loss term.

C. Our Training Strategy

For learning a person-specific model, we adopt a two-step training process. We train the first and second steps on the person's unoccluded and occluded face images, respectively. In the first step, we freeze the attention module and only train the encoder-decoder to reconstruct the unoccluded images. We start with unsupervised training of only encoder-decoder on publicly available state-of-the-art face datasets such as VGGFace [4] and AffectNet [12] to leverage the inherent knowledge about the face structure. It enables the model to grasp knowledge about the basic facial structure and features such as eyes, nose, mouth, etc. Thus this step is common for all users. We then perform finetuning on users' images with a wide range of pose, expression, and appearance variations. This step helps the model learn the exact geometry of the user's face. We unfreeze our attention module in the second step and train the entire architecture on occluded images. It can be considered a self-supervised learning approach, where the input is an occluded image, and the target image is its corresponding unoccluded image. We, therefore, minimize the loss between the reconstructed face image and the unoccluded ground truth image. It enables the attention module to learn to retain the high-frequency details from the visible part of the occluded input image while performing a soft fusion with the generated image. Thus, our two-step training strategy yields superior de-occlusion with no explicit boundaries between occluded and visible regions of the reconstructed image.

IV. Experiments and Results

A. Dataset

Our method uses monocular RGB video sequences. Hence, we captured various human subjects (around 20) in different appearances at a 1280 $\times$ 720 pixels resolution with a 30 FPS frame rate from a mobile phone camera.

We collected 4-5 video sequences for each user, each of a length of around 1-2 min, i.e., approximately 8k-9k frames in total. Frames are cropped and scaled to 256 × 256. We use mutually exclusive sets of these video sequences from the same subject in different appearances to train and evaluate our user-specific model. The subjects were asked to engage in a day-to-day conversation and demonstrate variations in head poses.

It is important to note that this data is captured without any occlusion to the eye region to create ground-truth data for training and evaluation purposes. Thus, we add a synthetic mask that simulates an HMD for each video frame around the eye region. The placement of this binary mask is guided by the facial landmarks and placed over the eye region to occlude the face, which yields synthetic mask data with ground-truth. During inference on real-world HMD occlusions, we first detect the HMD/smart-glass in the input video frames and replace it with a binary mask depicting the region that needs inpainting.

B. Implementation Details

For step one of training strategy (see Section III-C), we train the encoder-decoder architecture on unoccluded images with three loss functions (i.e., Equations 2, 3, 4) in a stage-wise manner, with each loss term being added in the training objective with every stage. We choose a batch size of 50 and an input resolution of 256 × 256. We train the network for 300, 100, and 300 epochs, respectively, for each loss function's incremental addition. For step two, we similarly train the encoder-decoder architecture with an additional attention module with mask loss (Equation 5) on occluded images for 300, 100, and 200 epochs, respectively, for each of the incremental addition of loss functions. We use the Adam optimizer [9] with a constant learning rate of 0.00002. We use $\lambda_{rec} = 1, \lambda_{adv} = 0.25, \lambda_{ssim} = 60$ and $\lambda_{mask} = 1$.

C. Evaluation Protocols

We choose SSIM (Structural Similarity Index Measure [25], PSNR (Peak Signal-to-Noise Ratio) [27] and LPIPS (Learned Perceptual Image Patch Similarity) [26] as our evaluation metrics for quantitative comparisons. For SSIM and PSNR, higher the value better the reconstruction quality, and for LPIPS, lower the value better the perceptual quality.

D. Quantitative & Qualitative Results

For qualitative evaluation, we evaluate our method with real occlusions, such as smart-glass, widely used in VR/AR applications. We overlay the area surrounding the smart glass with a synthetic mask generated (Section IV-A).

Figure 5 shows that our approach produces naturally-looking de-occluded faces that are semantically consistent with other frames in the sequence. In contrast, other state-of-the-art image inpainting methods like DeepFillv2 [21], LaFIn [23], EdgeConnect [24], when fine-tuned on images of the same user in different expressions, poses, and appearances, generates poor reconstruction results. As shown in the red box, these methods have a noticeable discrepancy between the left and right eyes. There are overlapping artifacts around the eye region indicated by a blue box. The yellow patch shows the inconsistency in skin color between the hallucinated and the rest of the face. This visual comparison strongly supports our idea of using a person-specific training approach rather than the generalized method since they do not guarantee to preserve identity and other high-frequency details such as appearance, pose, and expressions across frames. We also report de-occlusion results using our method in varying expressions and head poses in Figure 6 to verify the generalizability of the proposed model.

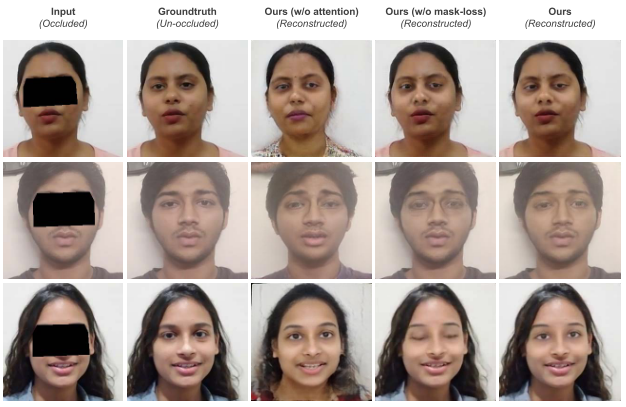

Figure 4. Visual results on unseen appearance demonstrating the effect of using attention and mask-loss. From left to right, third column shows results of our method without attention and mask-loss, fourth column shows results with only attention and fifth column shows results with both attention and mask-loss.

Table I reports the quantitative evaluation indicating the benefit of our approach for unseen appearances. It is important to note that we computed these quantitative results on data with a synthetic HMD mask to have a ground-truth to compare with. We can observe that our method achieves superior results in terms of all evaluation metrics. Though our method seems to perform only marginally better than LaFIn and EgdeConnect in terms of SSIM and LPIPS measures, there is a significant difference in terms of qualitative results. DeepFillv2 fails poorly in hallucinating the missing region, thus reporting higher LPIPS and lower

Table I
QUANTITATIVE COMPARISON WITH OTHER METHODS ON FACE RECONSTRUCTION.

Method	SSIM↑	PSNR↑	LPIPS↓
LaFIn [23]	0.914	23.693	0.0601
EdgeConnect [24]	0.908	23.10	0.0689
DeepFillv2 [21]	0.845	19.693	0.117
Ours (w/o attention)	0.706	19.627	0.176
Ours	0.938	30.59	0.029

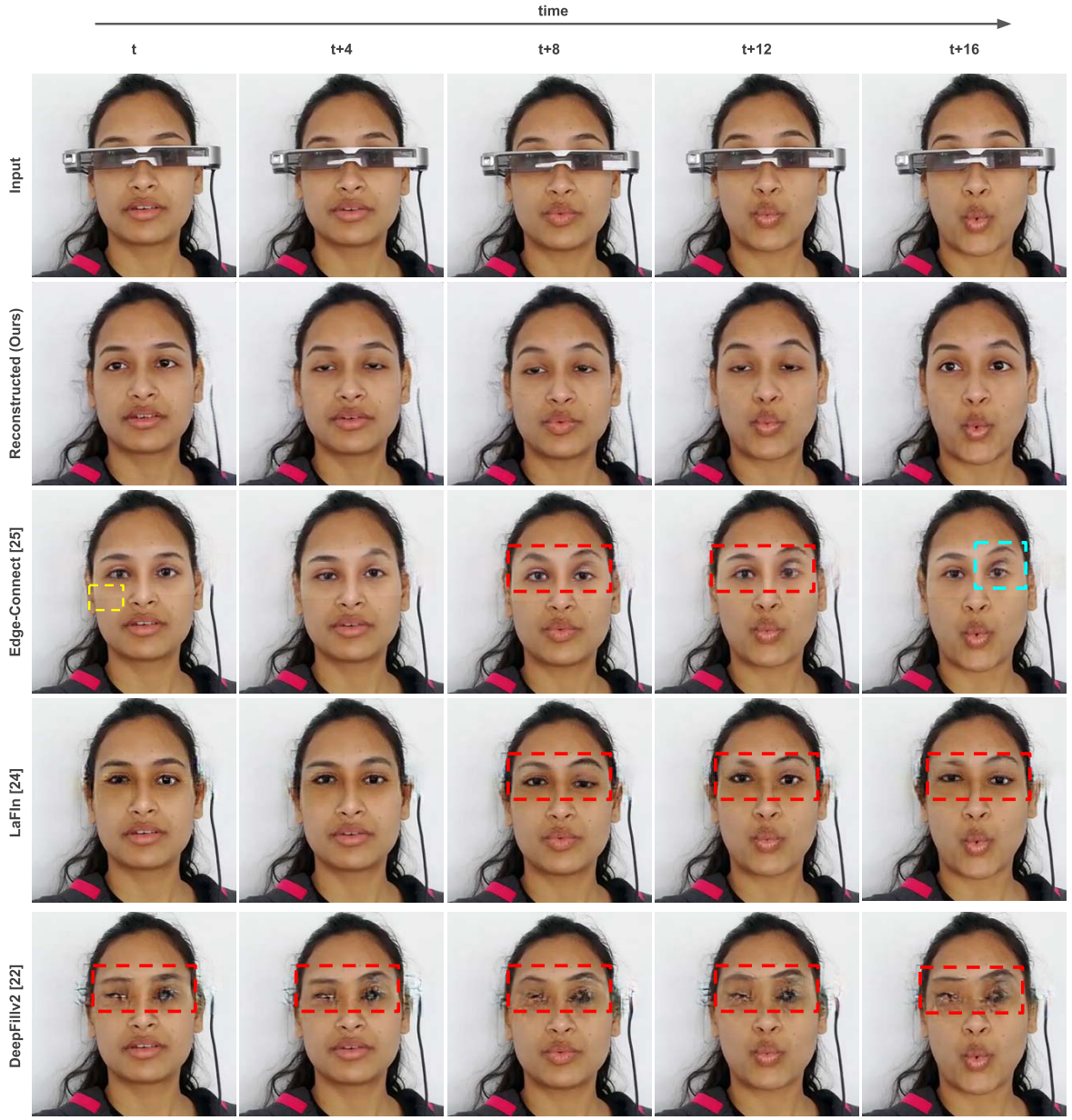

Figure 5. Qualitative comparison with SOTA inpainting methods on real-world occlusion (smart-glass). Zoom in for better details.

SSIM, PSNR compared to our method.

E. Ablation Study

We also did an ablation study on the various loss functions used in our method. We observed that training only with L_{rec} generates a blurry image, whereas the addition of L_{adv} introduces more sharpness. Finally, L_{ssim} and L_{mask} make it more consistent with facial features in the original image. Table II reports quantitative evaluation indicating incremental importance of all loss functions in our formulation.

Figure 4 qualitatively shows the effect of adding attention and mask-loss into the network. We use unseen test examples with different appearances (not part of the training set). Without attention, we can observe that the model cannot

Table II
ABLATION STUDY ON DIFFERENT LOSS FUNCTIONS.

Method	SSIM↑	PSNR↑	LPIPS↓
Ours (L_{rec})	0.932	30.28	0.067
Ours (L_{rec}+L_{adv})	0.916	29.37	0.042
Ours (L_{rec}+L_{adv}+L_{ssim})	0.936	30.37	0.031
Ours (L_{rec}+L_{adv}+L_{ssim}+L_{mask})	0.938	30.59	0.029

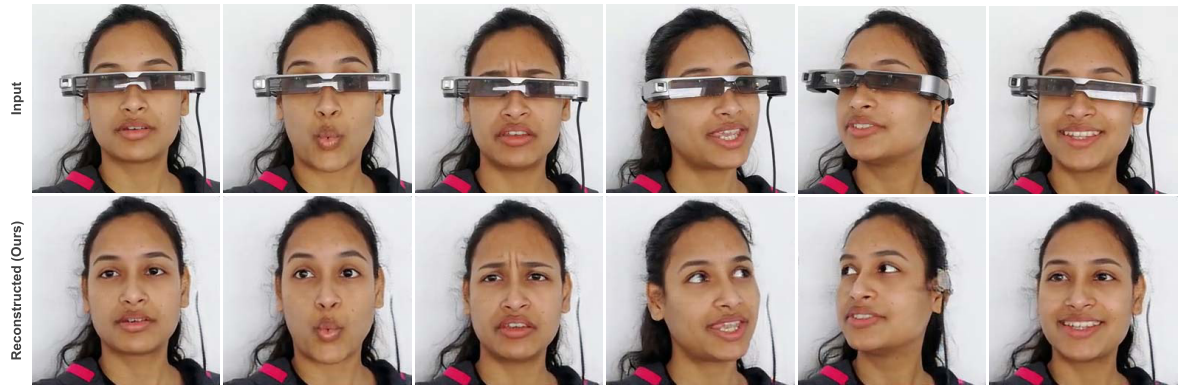

Figure 6. De-occlusion results using our method with large variations in head poses and expressions.

Table III
ABLATION STUDY ON DIFFERENT DIMENSIONALITY OF Z-VECTOR.

#Dims.	SSIM↑	PSNR↑	LPIPS↓
99	0.918	29.025	0.042
256	0.938	30.59	0.029
512	0.935	29.12	0.031

capture the user's appearance and background details as it tries to hallucinate it from the training examples, whereas introducing the attention into the network allows the model to use the high-frequency information from the input image. Furthermore, adding the mask loss introduces more consistency in the hallucination of the masked region. As can be observed in row 3 of Figure 4, introducing the mask allows the eyes to be open as it is with the ground truth face image. Table III reports an ablation study on the dimensionality of z-vector. We achieved better results with $d = 256$.

V. Application to Hybrid Telepresence System

Recent works on face video animation, such as [16], [19], demonstrate that by just using sparse landmarks, a face image can be animated reasonably well from a reference image and show its application in low-bandwidth environments. We can easily integrate their method in our setup by first de-occluding the HMD followed by extracting reliable sparse landmarks for facial animation as shown in Figure 7. Thus, we can generate a consistent 2D video feed from the input occluded video feed. Moreover, this animated face can also be used for per-frame 3D face reconstruction tasks [22] and fed to other VR teleconferencing users wearing a VR headset (as shown in Figure 8). Hence, our method allows VR and non-VR users to share a similar experience in a single hybrid teleconferencing application.

VI. Discussion

As shown earlier, our proposed approach promises to give superior results, both qualitatively and quantitatively. We also notice that if HMD occlusion is more than 70 percent

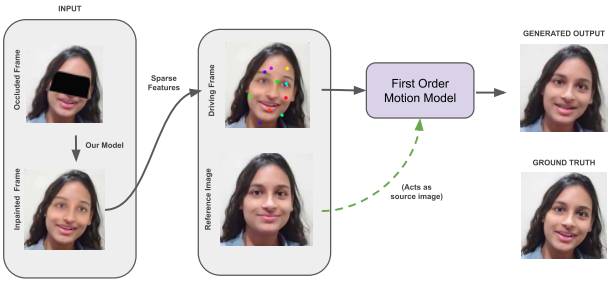

Figure 7. 2D de-occlusion using our method followed by the facial animation using FOMM [16].

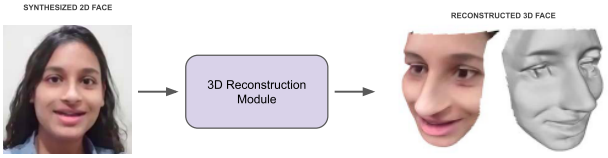

Figure 8. 3D reconstruction of de-occluded frame using DF2Net [22].

of the face, it becomes difficult for any existing methods to reconstruct plausible faces for varying expressions and head poses. Hence, these models can reconstruct canonical faces, which might not always be the case in telepresence systems.

In the scope of this work, our model does not explicitly handle eye movements since we are not providing any strong priors such as landmarks. Thus, it might not be able to capture accurate eyelid movement during blinks. However, we can easily incorporate eye tracking and gaze information for further refining our results. As future work, we would like to focus on performing HMD de-occlusion by leveraging temporal information across consecutive frames. Using modern HMD devices, we can give extra supervision about the eye information to the network to produce stable reconstruction in the eye region that is consistent across frames.

VII. Conclusions

We proposed to learn a personalized model for face de-occlusion, particularly as an application in VR teleconferencing, where the face is partially occluded due to HMD. We formulate the face de-occlusion problem as an image inpainting task. Our proposed attention enabled encoder-decoder network takes an HMD occluded face as input and completes missing facial features, particularly the eye region. The experiments show that our method works reasonably well with the same person wearing different clothes, facial appearances, poses, and expressions. Experiments show that our proposed method reports superior qualitative and quantitative results over state-of-the-art methods.

References

[1] V. Blanz and T. Vetter. A Morphable Model For The Synthesis Of 3D Faces. In Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH'99.

[2] K. W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition. Computer vision and image understanding, 2006.

[3] B. Browatzki and C. Wallraven. 3FabRec: Fast Few-shot Face alignment by Reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition '20.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: A Dataset for Recognising Faces across Pose and Age. IEEE International Conference on Automatic Face and Gesture Recognition, 2018.

[5] M. Elgharib, M. Mendiratta, J. Thies, M. Niessner, H.-P. Seidel, A. Tewari, V. Golyanik, and C. Theobalt. Egocentric videoconferencing. ACM Transactions on Graphics (TOG) '20.

[6] G. Gafni, J. Thies, M. Zollhofer, and M. Nießner. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks, 2014.

[8] P. Isola, J.Y. Zhu, T. Zhou, and A. A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.

[9] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. arXiv:1412.6980, 2014.

[10] A. Kubade, D. Patel, A. Sharma, and K. Rajan. AFN: Attentional Feedback Network based 3D Terrain Super-Resolution. In Proceedings of the Asian Conference on Computer Vision '20.

[11] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh. Deep appearance models for face rendering. ACM Transactions on Graphics, 2018.

[12] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 2017.

[13] N. Numan, F. ter Haar, and P. Cesar. Generative RGB-D Face Completion for Head-Mounted Display Removal. In 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW).

[14] A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, 2016.

[15] Kelvin Xu. Jimmy Ba, R. Kiros, K.Cho, A.Courville, R. Salakhutdinov, R.Zemel, Y.Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. https://arxiv. org/abs/1502.03044 v3.

[16] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. First Order Motion Model for Image Animation. Advances in Neural Information Processing Systems, 2019.

[17] J. Thies, M. Zollofer, M. Stamminger, C. Theobalt, and M. Nießner. FaceVR: Real-Time Facial Reenactment and Eye Gaze Control in Virtual Reality. arXiv:1610.03151 '16.

[18] M. Wang, X. Wen, and S.M. Hu. Faithful Face Image Completion for HMD Occlusion Removal. In 2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct.

[19] T.C. Wang, A. Mallya, and M.Y. Liu. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[20] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Free-Form Image Inpainting with Gated Convolution. arXiv:1806.03589, 2018.

[21] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative Image Inpainting with Contextual Attention. arXiv:1801.07892, 2018.

[22] X. Zeng, X. Peng, and Y. Qiao. DF2Net: A Dense-Fine-Finer Network for Detailed 3D Face Reconstruction. ICCV 2019.

[23] Y. Yang, X. Guo, J. Ma, L. Ma, and H. Ling. LaFIn: Generative Landmark Guided Face Inpainting, 2019.

[24] K. Nazeri and E. Ng and T.Joseph and F. Qureshi and M. Ebrahimi. EdgeConnect: Structure Guided Image Inpainting using Edge Prediction, ICCV Workshops 2019.

[25] Wang, Z. , Bovik, A. C. , Sheikh, H. R. and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing, 2004.

[26] R.Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, CVPR 2018.

[27] A. Horé and D. Ziou. Image Quality Metrics: PSNR vs. SSIM, ICPR, 2010.