Supervision by Landmarks: An Enhanced Facial De-occlusion Network for VR-based Applications

Surabhi Gupta, Sai Sagar Jinka, Avinash Sharma, and Anoop Namboodiri

Center for Visual Information Technology
International Institute of Information Technology Hyderabad (India)
{surabhi.gupta, jinka.sagar}@research.iiit.ac.in
{asharma, anoop}@iiit.ac.in

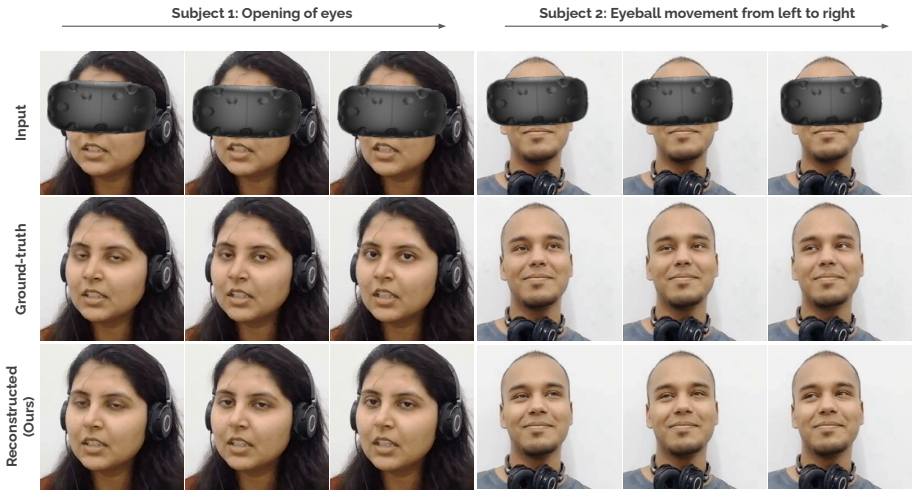

Fig. 1. Photo-realistic results generated by our proposed facial de-occlusion network, targeting complex eye motions.

Abstract. Face possesses a rich spatial structure that can provide valuable cues to guide various face-related tasks. The eyes are considered an important socio-visual cue for effective communication. They are an integral feature of facial expressions as they are an important aspect of interpersonal communication. However, virtual reality headsets occlude a significant portion of the face and restrict the visibility of certain facial features, particularly the eye region. Reproducing this region with realistic content and handling complex eye movements such as blinks is challenging. Previous facial inpainting methods are not capable enough to capture subtle eye movements. In view of this, we propose a working solution to refine the reconstructions, particularly around the eye region, by leveraging inherent eye structure. We introduce spatial supervision and a novel landmark predictor module to regularize per-frame reconstructions obtained from an existing image-based facial de-occlusion network. Experiments verify the usefulness

of our approach in enhancing the quality of reconstructions to capture subtle eye movements.

Keywords: face image inpainting, landmark guided facial de-occlusion, HMD removal, virtual reality, eye consistency

1 Introduction

Social telepresence and interaction are essential for human survival. Since globalization, there has been a considerable increase in users interacting remotely, which has witnessed a tremendous surge during the Covid-19 pandemic. Traditional video conferencing platforms such as Microsoft Teams, WhatsApp, etc., gained immense popularity during the pandemic. However, they lack immersiveness and compromise realism that impacts the user's experience, which is undesirable. With the integration of virtual reality in a communication platform, the current technologies have witnessed a breakthrough in enhancing user experience with a sense of heightened social existence and interaction. Faces convey vital socio-visual cues that are important for effective communication. However, one of the major challenges with virtual reality, such as HMDs, is the occlusion it cause over the face when wearing these devices. These devices occlude almost 30-40 percent of the face, obscuring essential social cues, particularly the eye region, which hinders the user's experience. Several approaches have been proposed in the literature to tackle this problem, but none of them produces photorealistic results that could be integrated into hybrid telepresence systems.

Existing face image inpainting approaches often suffer from incoherency in generating smooth reconstructions when applied to video frames. This incoherency is highly noticeable in the eye region, which is undesirable. It is generally visible as jittering in eyelids in successive video frames. Specific eye movements, such as blinking, are usually involuntary act in humans that is natural and unavoidable. Thus, it is important to retain this characteristic for effective communication. Synthesizing eyes, including iris and eyelid reconstruction with appropriate eye gaze, have been attempted before using 3D model-based approaches. Nonetheless, they require high-quality data and incur expensive training costs. [6] and [17] are such examples of 3D models based approaches to HMD de-occlusion. However, they work only with frontal face images and fail in cases of extreme head-poses. Another set of works like [14], [15] use an inpainting approach to correct and animate the eye gaze of high-resolution, unconstrained portrait face images. Since these methods have not been validated on videos, they fail to generate consistent eye motions across frames.

Interestingly, one of the biggest advantages when dealing with digital face images in computer vision is its rich spatial structure. In digital images, this structure is generally represented in the form of 2D/3D coordinates, heatmaps, and edges and is provided as an auxiliary input to the network. Many works in the literature have exploited these spatial constraints for achieving better quality reconstruction in face inpainting and generation tasks. Recently, [11] proposed image-to-face video inpainting using spatio-temporal nested gan architecture. They used 3D residual blocks to capture inter-frame dependencies. The authors showed that conditioning facial inpainting on landmarks

yielded stable reconstructions. Nonetheless, it is only validated with a specific type of circular mask that covers the eye, nose, and mouth. [12] is another face image inpainting method that is guided using facial landmarks. Personalized facial de-occlusion networks such as [3] have been proposed in the literature to generate plausible reconstructions. However, they are not controllable and thus cannot handle eye movements. Thus, we tackle this problem using an image generation/image synthesis approach applied to faces using additional information that is easily accessible using modern HMD devices with eye-tracking capabilities.

This work aims to generate high-quality facial reconstructions in and around the eye region with consistent eye motions in the presence of occluders such as HMDs. Our primary focus is to handle instability during eye movements that are noticeable mainly around the eye region. For this, we leverage the spatial property of faces, i.e., facial landmarks, to guide the model to synthesize the eye region with minimal artifacts that look realistic and plausible. Figure 1 presents high-quality and photo-realistic results produced by our method showing the efficacy of our approach of using spatial supervision to control complex eye motions such as eye blinks and rolling of eyeballs.

To summarize, we make the following contributions:

1. We propose a potential solution to refine the reconstructions in the eye region.
2. To achieve this, we leverage the spatial constraints such as landmarks to improve upon consistency in the eye region by feeding eye landmarks heatmaps as an auxiliary input to the network along with occluded face image.
3. To further improve the fidelity of the reconstruction, we use an additional loss function to regularize the training based on the landmarks.

2 Related Work

To see what is not present in the image is one of the most exciting yet challenging tasks in computer vision. We often refer to it as image restoration/image inpainting in the digital domain. It has applications in medical image processing, watermark removal, restoring old photographs, and object removal. Inpainting has been an active research topic for many years, and several works have been proposed in the literature. Recently, this area has seen tremendous interest in image synthesis/image completion in AR/VR. This section will discuss the most relevant existing works in detail.

2.1 Facial De-occlusion and HMD Removal Methods

De-occluding face images in the presence of large occluders such as HMDs is highly an ill-posed problem. Several works such as [6], [9] have been proposed in the literature to address this issue. However, none promises to provide usable results in practice as these have only been validated for frontal face images with rectangular masks. Since they use an additional reference image of the person, they fail in cases of different pose variations between occluded and reference images. Recently, [3] presented an approach to tackle the problem of facial de-occlusion by training a person-specific model in VR settings. It generates plausible and natural-looking reconstructions but might fail to maintain

smooth eye movements across consecutive frames. To address this issue, we can use extra information provided by modern HMD devices equipped with eye-tracking to generate consistent eye motions.

2.2 Structure-guided Image Inpainting

Figuring out missing regions without any prior information is a difficult task. Many prior works have successfully used landmarks for the task of face generation and synthesis. Previous image inpainting methods, such as [5], [12] use edges, landmarks, and other structural information as an auxiliary input to guide the reconstructions. This extra supervision has proven effective in helping the model fill the missing region with appropriate content. However, these are image-based approaches and might not guarantee to generate consistent results across frames. Thus, we cannot directly use these methods to generate smooth reconstructions, particularly in the eye region.

3 Proposed Method

3.1 The Architecture

We built upon the architecture proposed in [3] and used an attention-enabled encoder-decoder architecture followed by a novel Landmark Heatmap Predictor (LHP) module that acts as a regularizer to enhance the reconstruction in and around the eye region. We train this network in an end-to-end fashion in two stages using a dedicated loss function. [3] is an existing facial de-occlusion network, and we consider it a baseline.

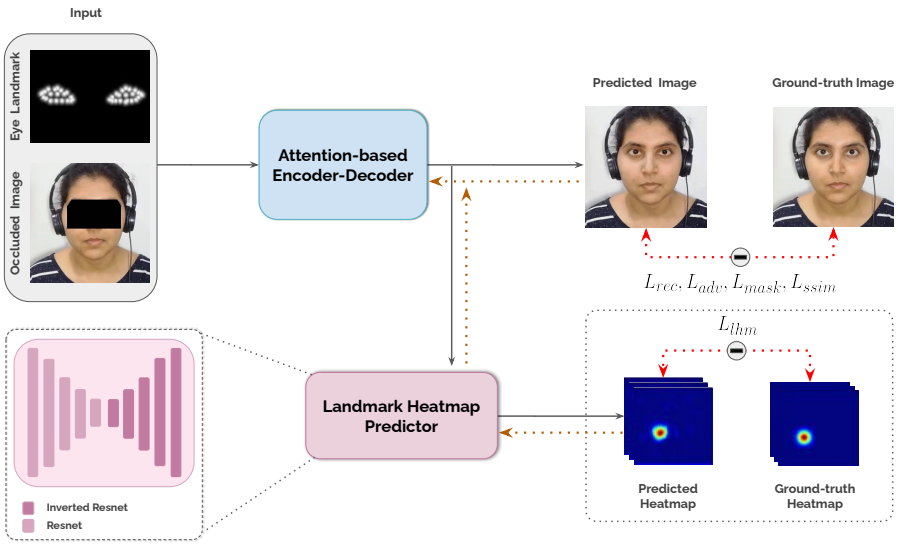

Fig. 2. Illustration of our proposed architecture.

Attention-based encoder-decoder: We utilize an encoder-decoder architecture with an attention module to inpaint the missing regions of the face, particularly the eye region. The primary function of the attention module is to focus on reconstructing this region with high-frequency details such as hair, facial accessories, and appearances. It also helps the model to generalize to unseen and novel appearances, hairstyles, etc. The encoder-decoder comprises ResNet and Inverted ResNet layers, with a bottleneck layer of 99 dimensions. For Inverted ResNet, the first convolution in the ResNet block is replaced by a 4×4 deconv layer. The attention module is composed of four convolution layers: $Conv(4 * m, 3)$, $Conv(4 * m, 3)$, $Conv(8 * m, 3)$ and $Conv(2 * m, 3)$, 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. Given an occluded face image as an input X_{occ}, this network hallucinates the missing region in order to reconstruct the generated unoccluded image, X_{rec} against the ground truth unoccluded image, X_{gt}.

Landmark heatmap predictor module: We employ another encoder-decoder network to refine the reconstruction around the eye region. The primary aim of this network is to predict the eye landmark heatmap of the reconstructed image, based on which we can regularize the final reconstructed image using a loss function. This landmark heatmap predictor network is composed of ResNet and Inverted ResNet layers. The input to this network is the reconstructed image, X_{recon} produced from the attention-based encoder-decoder network. The output is a 42-channeled landmark heatmap, denoted by $LHP(X_{rec})$, where each channel corresponds to one of 42 eye landmarks. Figure 2 illustrates an overview of the proposed pipeline.

3.2 Spatial Supervision using Landmarks

Per-frame predictions from traditional image-based facial de-occlusion network such as [3] suffer from temporal discontinuity and flickering, especially in eyelids. Therefore, to stabilize the eye movements, we leverage eye landmarks as an auxiliary input to guide smooth reconstructions in the eyelids that are much more realistic and consistent. This supervision helps the model preserve the structure of the eyelids. For better and enriched representations, we prefer 2D heatmaps over 2D coordinates. Each landmark is represented by a separate heatmap, interpreted as a grayscale image. We convolve all heatmaps to a single-channel grayscale image which is then concatenated with the occluded RGB input image in the channel dimension. This is further fed to the attention-based encoder-decoder to generate plausible and stable reconstructions.

3.3 Loss Functions

The primary goal of this pipeline is to generate plausible facial inpainted reconstructions consistent with other frames in sequence while preserving the landmark structure of the eyes. To serve this purpose, we use the following loss functions:

The first loss function ensures that the generated reconstruction is in close proximity to the ground-truth unoccluded image. Thus, we formulate pixel-based $L1$ loss to penalize reconstruction errors.

$$L_{rec} = \|X_{gt} - X_{rec}\|_1 \tag{1}$$

However, using only reconstruction loss generates blurry reconstructions. Thus we adopt the architecture of the DCGAN discriminator [7] in the pipeline, denoted by D, to compute the adversarial loss that forces the encoder-decoder to reconstruct high-fidelity outputs by sharpening the blurred images.

$$L_{adv} = log(D(X_{gt})) + log(1 - D(X_{rec})) \tag{2}$$

To further stabilize the adversarial training, we use SSIM based structural similarity loss, as defined in [1], that helps to improve the alignment of high-frequency image elements.

$$L_{ssim} = SSIM(X_{rec}, X_{gt}) \tag{3}$$

In order to emphasize the quality of reconstruction in the masked region, i.e., invalid pixels, we use a mask-based loss function. Here, we use the binary mask image as additional supervision to the network and input image while training. This helps mitigate the blinking artifacts around the eye region for stable reconstructions.

$$L_{mask} = \|I_{mask} \odot X_{gt} - I_{mask} \odot X_{rec}\|_1 \tag{4}$$

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.

Apart from providing a landmark heatmap along with occluded input, we also regularize the reconstructions based on landmarks using a loss function. To prevent irregularities in the eye region and preserve eyelid shape, we utilize landmark heatmap prediction loss that regularizes the inpainted reconstructions based on predicted eye landmark heatmaps, $LHP(X_{rec})$ and ground-truth eye landmark heatmaps, H. Here, for each landmark $l_i \in R^2$, H_i consists of a 2D normal distribution centered at li and a standard deviation of σ.

$$L_{lhm} = \|H - LHP(X_{rec})\|_2 \tag{5}$$

Thus, the final training objective loss function can be written as,

$$L_{final} = \lambda_{rec} * L_{rec} + \lambda_{adv} * L_{adv} + \lambda_{ssim} * L_{ssim} + \lambda_{mask} * L_{mask} + \lambda_{lhm} * L_{lhm} \tag{6}$$

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

4 Experiments and Results

4.1 Dataset and Training Settings

Dataset preparation: We train the network on different face video sequences for multiple identities. We train a person-specific model for every identity on 4-5 sequences captured in various appearances, including apparel, hairstyle, facial accessories, and different head poses. Videos are recorded at a resolution of 1280 x 720 at 30 fps using a regular smart-phone and then cropped to 256 x 256 for training. Note that there is

no overlap between the training and test set. To test the ability of our model to generalize to novel appearances, we validate it with completely unseen videos that are not seen during the training process. The dataset is available here. For the provision of spatial supervision, we use Mediapipe [2] to detect and localize 42 landmarks around the eye, including iris landmarks. As discussed in Section 3.2, we create a heatmap for every landmark coordinate. Since we extract pseudo landmarks directly from unoccluded ground truth that is inherently spatially aligned with the occluded face, we directly append landmark heatmaps with the occluded input image without any further processing.

Inference with real occlusion: The eye information might not be directly accessible during inference when wearing regular virtual reality headsets. Fortunately, modern devices allow eye tracking using IR cameras mounted inside headsets. We can extract this information from eye images captured using these cameras. Unfortunately, the images captured by these cameras are not aligned with the face image. Hence, we need to calibrate both the eye and face camera as proposed in [8], [17] to align the eye images with the face image coordinate system. However, due to the unavailability of these headsets, we opt for pseudo landmarks extracted from ground-truth images to provide supervision to the model. As discussed, we extract these landmarks using the Mediapipe [2] face landmark detector. It is to be noted that these landmarks do not adhere well to an anatomically defined point across every video frame and thus have local noise in them generated due to the inaccuracy of the facial landmark detector.

Training strategy: We follow a similar two-stage training strategy proposed in [3]. In the first stage, we only train the encoder-decoder network without an attention module on unoccluded images along with their corresponding eye landmark images of the person using the first three losses aforementioned in Section 3.3, each added incrementally after 400, 100, and 300 epochs, respectively. In the second stage, we finetune the same encoder-decoder with the attention module and the landmark prediction module on occluded images of the same person and their corresponding eye landmark images using two additional loss functions. We use the same three losses as the first stage. Apart from this, we also use a landmark heatmap prediction loss to regularize the reconstructions generated from the attention-based encoder-decoder network and a mask-based loss to minimize reconstruction errors in the masked region. We use $\lambda_{rec} = 1, \lambda_{adv} = 0.25, \lambda_{ssim} = 60, \lambda_{lhm} = 1$ and $\lambda_{mask} = 1$.

4.2 Results

In this section, we present the results of our method and discuss its superiority over existing approaches. We first compare the visual quality of the reconstruction generated by our method with popular state-of-the-art inpainting methods, followed by a quantitative analysis using standard evaluation metrics. To further validate the efficacy of our approach, we also report an ablation study conducted in the scope of this work.

Qualitative comparison: For visual comparisons, we evaluate our method against various image inpainting methods across 20 subjects. Results highlighted in Figure 4, 6, 7

show that the reconstructions generated using our method are visually pleasing and consistent across frames compared to other inpainting methods. Reconstructions generated using our approach, as shown in row (C) of Figure 4 show the significance of landmark supervision and regularization loss in capturing eye movements such as blinks. However, predictions using other approaches are often incoherent across frames. As visible in row (F), DeepFillv2 [13] fails poorly to generate plausible reconstruction in the eye region. LaFIn [12] and Edge-Connect [5] generate superior reconstructions compared to DeepFillv2, however, it cannot handle eye movements. Besides, there is a noticeable discrepancy in the left and right eyes that looks unnatural. Baseline [3] produces naturally-looking reconstructions but cannot handle eye blinks. For better comparison, refer to the supplementary video. In Figure 3, we also show the reconstruction error ($l2$ error) between the results generated by different image inpainting methods and the ground truth for better justification. Please refer to the supplemental video.

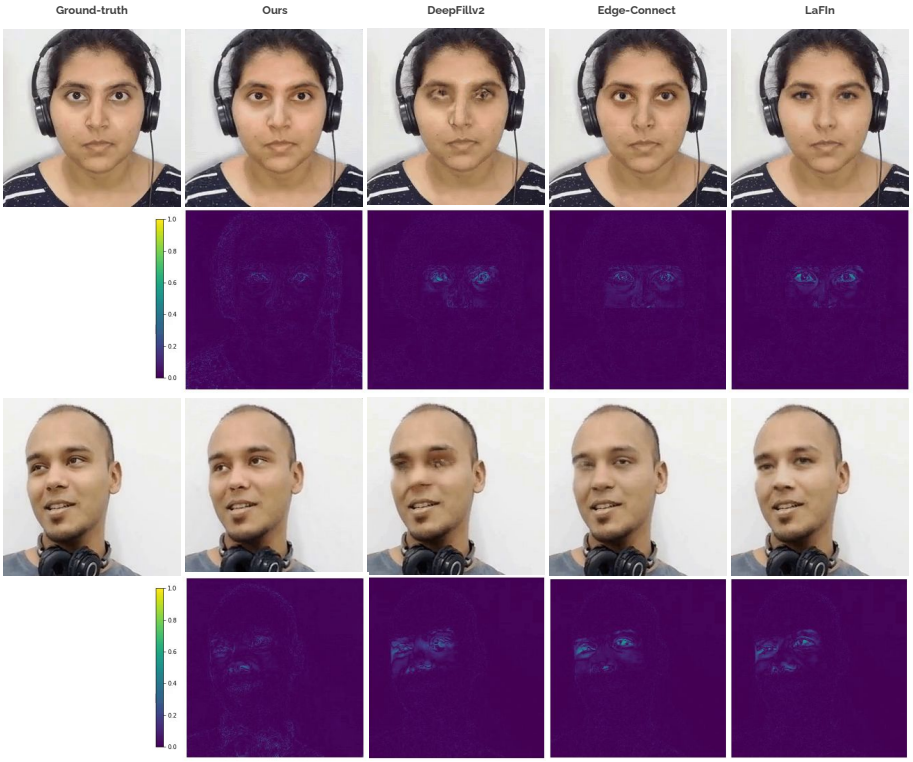

Fig. 3. Qualitative result that showing the reconstruction error ($l2$ error) between the results generated by different image inpainting methods and the ground truth.

Fig. 4. Qualitative comparison with SOTA image inpainting methods. From row (A-G) are Occluded (input), Original (ground-truth), Ours, Baseline [3], Edge-connect [5], DeepFillv2 [13] and LaFIn [12] respectively. From left to right are consecutive frames of unseen testing video.

Quantitative comparison: To quantify the quality of reconstructions, we use standard image quality metrics such as SSIM [10], PSNR [4], and LPIPS [16]. For SSIM and PSNR, a higher value indicates better reconstruction quality and vice-versa. Similarly, for LPIPS, a lower value indicates better perceptual quality and vice-versa. Table 1 shows the quantitative comparison of our proposed method with other state-of-the-art face inpainting methods. As reported, our method **(in bold)** performs better in all evaluation metrics than other methods such as Edge-connect [5], DeepFillv2 [13], Baseline [3] and LaFIn [12].

Method	SSIM↑	PSNR↑	LPIPS↓
LaFIn [12]	0.914	23.693	0.0601
EdgeConnect [5]	0.908	23.10	0.0689
DeepFillv2 [13]	0.845	19.693	0.117
Baseline[3]	0.918	29.025	0.042
Ours	**0.949**	**31.417**	**0.0235**

Table 1. Quantitative comparision of our method with other state-of-the-art image inpainting methods.

5 Ablation Studies

We perform several ablation studies to understand the various aspects of our model. We first analyze the effect of providing spatial supervision to the model in enhancing reconstruction quality, both qualitatively and quantitatively. As depicted in Figure 5 and 8, our model with landmarks produces aesthetically pleasing eyes and preserves eyelid shape in contrast to the one without landmarks supervision. It is due to the guidance provided by the landmarks that helps the model enforce consistency in eye movements, including the opening and closing of eyes. However, it is to be noted that this does not ensure eye movements are temporally coherent. Secondly, we show the effect of using a regularizing loss function based on landmarks heatmap to penalize the errors caused by the model. Table 2 reports the positive impact of using eye landmarks and landmark-based loss function in guiding the reconstruction in the eye region.

Method	SSIM↑	PSNR↑	LPIPS↓
Baseline[3]	0.918	29.025	0.042
Ours (with LHM)	0.926	29.272	0.0418
Ours (with LHM + L_{lhm}**)**	**0.949**	**31.417**	**0.0235**

Table 2. Ablation study showing the significance of using landmark supervision on the reconstruction quality. Here, LHM is the auxiliary landmark heatmap provided along with the input image and L_{lhm} is the regularizing loss function.

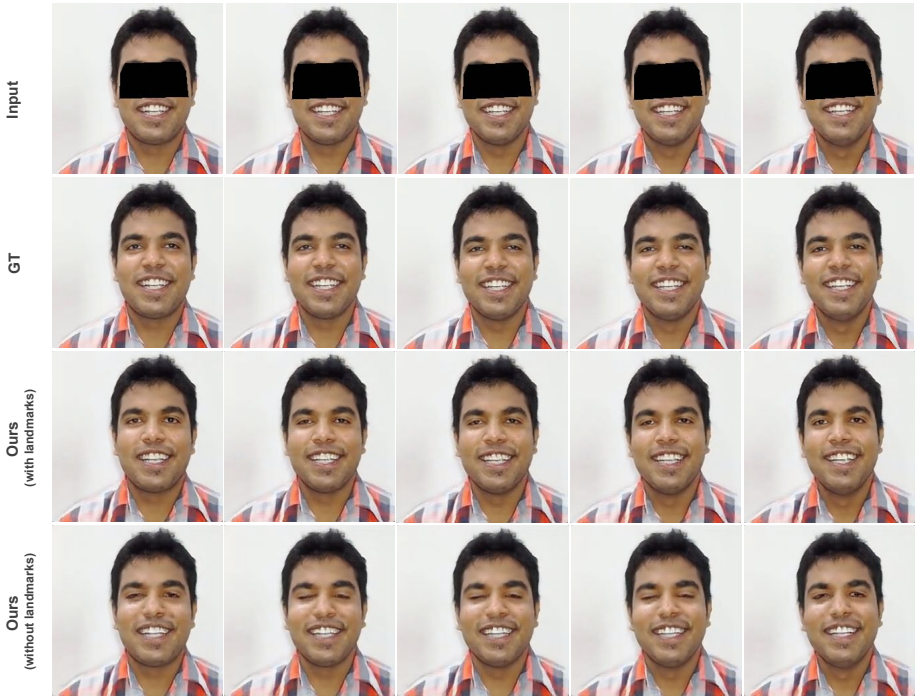

Fig. 5. Testing results showing the effect of using landmarks as auxillary input to the network. From row (1-4) are occluded (input), original (ground-truth), results with and without landmarks respectively. From left to right is temporal continuously images of original 30fps videos.

6 Conclusion

We present this work as an enhancement in existing facial de-occlusion networks by explicitly focusing on improving eye synthesis. We show that providing landmark information during the inpainting process can yield superior quality and photorealistic reconstructions, including the eye region. We discuss how this information can be retrieved: 1) by extracting pseudo landmarks from ground-truth images and 2) using modern HMD devices capable of tracking eye movements. To further enhance the generated output, we propose a landmark-based loss function that act as a regularizing term to improve reconstruction quality and helps capture subtle eye movements such as eye blinks. We conducted qualitative and quantitative analysis and reported superior results with other SOTA inpainting methods to justify the usefulness of our approach.

Fig. 6. Qualitative comparison with SOTA image inpainting methods. From row (A-G) are Occluded (input), Original (ground-truth), Ours, Baseline [3], Edge-connect [5], DeepFillv2 [13] and LaFIn [12] respectively. From left to right are consecutive frames of unseen testing video.

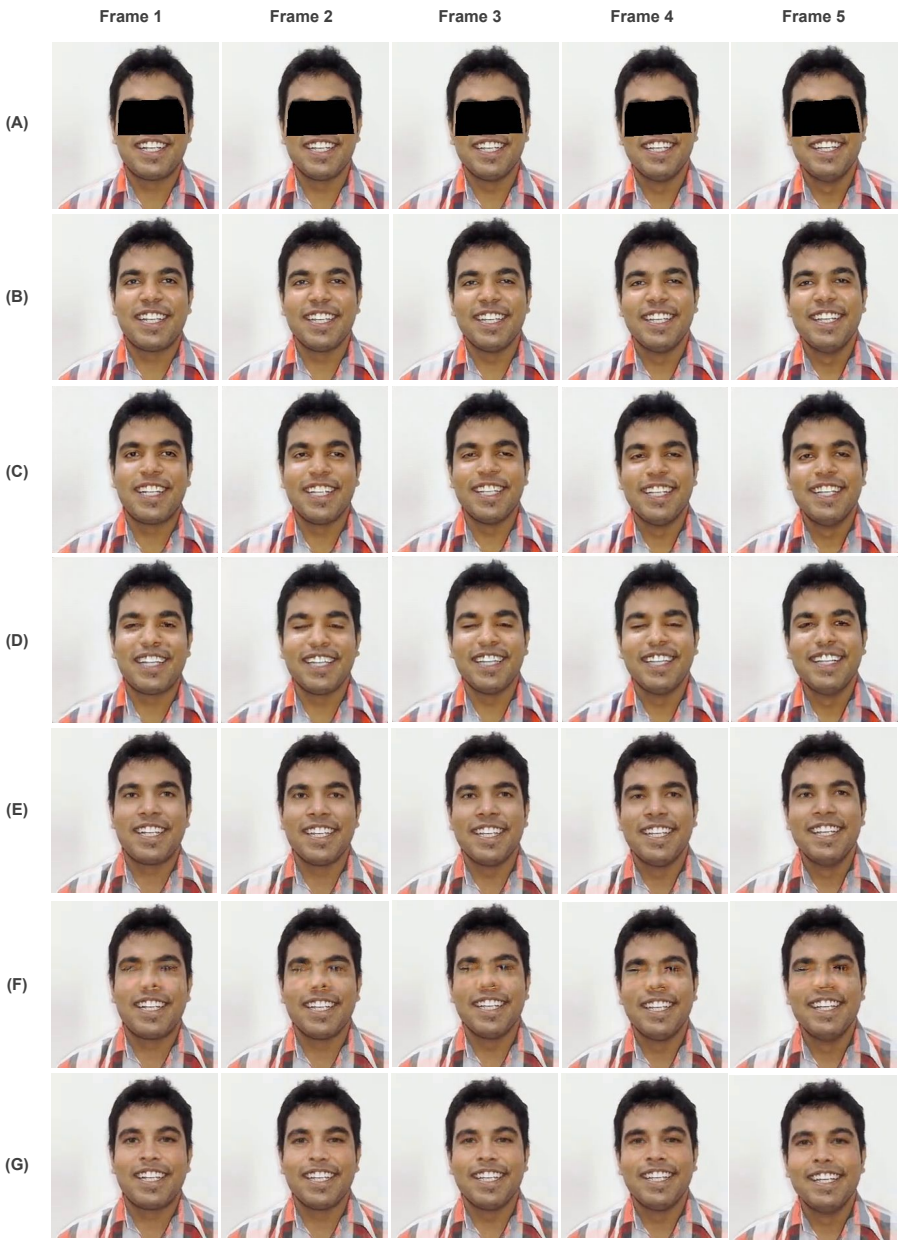

Fig. 7. Qualitative comparison with SOTA image inpainting methods. From row (A-G) are Occluded (input), Original (ground-truth), Ours, Baseline [3], Edge-connect [5], DeepFillv2 [13] and LaFIn [12] respectively. From left to right are consecutive frames of unseen testing video.

Fig. 8. Testing results showing the effect of using landmarks as auxillary input to the network. From row (1-4) are occluded (input), original (ground-truth), results with and without landmarks respectively. From left to right is temporal continuously images of original 30fps videos.

References

1. Browatzki, B., Wallraven, C.: 3fabrec: Fast few-shot face alignment by reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
2. Grishchenko, I., Ablavatski, A., Kartynnik, Y., Raveendran, K., Grundmann, M.: Attention mesh: High-fidelity face mesh prediction in real-time (2020)
3. Gupta, S., Shetty, A., Sharma, A.: Attention based occlusion removal for hybrid telepresence systems. In: 19th Conference on Robots and Vision (CRV) (2022)
4. Horé, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. ICPR (2010)
5. Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M.: Edgeconnect: Generative image inpainting with adversarial edge learning (2019)
6. Numan, N., ter Haar, F., Cesar, P.: 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). IEEE (2021)
7. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks (2016)
8. Thies, J., Zollöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: FaceVR: Real-Time Facial Reenactment and Eye Gaze Control in Virtual Reality (2016)
9. Wang, M., Wen, X., Hu, S.M.: Faithful face image completion for hmd occlusion removal. In: 2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). IEEE (2019)
10. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing (2004)
11. Wu, Y., Singh, V., Kapoor, A.: From image to video face inpainting: Spatial-temporal nested gan (stn-gan) for usability recovery. In: 2020 IEEE Winter Conference on Applications of Computer Vision (2020)
12. Yang, Y., Guo, X., Ma, J., Ma, L., Ling, H.: Lafin: Generative landmark guided face inpainting (2019)
13. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention (2018)
14. Zhang, J., Chen, J., Tang, H., Sangineto, E., Wu, P., Yan, Y., Sebe, N., Wang, W.: Unsupervised high-resolution portrait gaze correction and animation. IEEE Transactions on Image Processing (2022)
15. Zhang, J., Chen, J., Tang, H., Wang, W., Yan, Y., Sangineto, E., Sebe, N.: Dual in-painting model for unsupervised gaze correction and animation in the wild. In: ACM MM (2020)
16. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
17. Zhao, Y., Xu, Q., Chen, W., Du, C., Xing, J., Huang, X., Yang, R.: Mask-off: Synthesizing face images in the presence of head-mounted displays. In: 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR) (2019)