https://doi.org/10.2352/EI.2025.37.8.IMAGE-273

Frontal View Synthesis For Immersive Video Conferencing Using Dual-camera Capture and Frame Interpolation

Yezhi Shen[1], *Md Adnan Faisal Hossain*[1], *Weichen Xu*[1], *Qian Lin*[2], *and Fengqing Zhu*[1]

[1] *Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN*
[2] *HP Inc, Palo Alto, CA*

Abstract

In this paper, we propose a new solution for synthesizing frontal human images in video conferencing, aimed at enhancing immersive communication. Traditional methods such as center staging, gaze correction, and background replacement improve the user experience, but they do not fully address the issue of off-center camera placement. We introduce a system that utilizes two arbitrary cameras positioned on the top bezel of a display monitor to capture left and right images of the video participant. A facial landmark detection algorithm identifies key points on the participant's face, from which we estimate the head pose. A segmentation model is employed to remove the background, isolating the user. The core component of our method is a video frame interpolation technique that synthesizes a realistic frontal view of the participant by leveraging the two captured angles. This method not only enhances visual alignment between users but also maintains natural facial expressions and gaze direction, resulting in a more engaging and life-like video conferencing experience.

Introduction

In recent years, video communication has become an integral part of daily life, facilitating everything from personal interactions to professional teleconferencing. However, despite its widespread adoption, the experience of face-to-face communication remains elusive in most video-based interactions. This limitation stems primarily from the design of existing devices, which typically rely on a single camera. Such setups restrict the viewing angle to a fixed perspective, preventing participants from seeing each other in a natural, multi-dimensional way. As a result, the lack of spatial depth and dynamic visual engagement undermines the sense of presence and connection central to in-person communication. To improve the telecommunication experience, numerous efforts have been made to develop engaging video systems so that remote users across the internet could feel like they meet in the same physical space. The viewers are enabled to view the other participants at desired viewing angles other than the captured position.

Various 3D reconstruction based methods have proposed the utilization of depth sensors to generate the geometry of the video participants and rerender the model for the viewer at the required viewing angle [1]. These methods, however, largely depend on the expensive hardware to perform high-resolution captures and reconstructions, resulting in difficulty in popularizing. With the improvement of artificial intelligence, some methods propose to leverage the flexibility of machine learning models to remap the low-dimensional 2D information from a monocular RGB camera to 3D latent space [2], [3]. These methods are successful in reducing the cost of capture devices, but due to the heavy cost of computation, they require top-end graphics cards to perform real-time inference. A third group of methods looks for a middle ground, where they design to use multiple RGB cameras for the capture and perform novel view synthesis directly on the 2D image space to maintain a reasonable low computation cost [4], [5]. These methods, however, are prone to background distractions of large motions, which distorts the edges of synthesized faces.

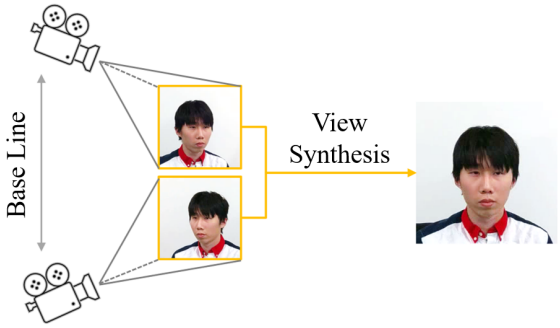


Figure 1: Illustration of our camera setup and frontal view synthesis method.

In this work, we introduce a frontal view synthesis method for immersive video conferencing based on video frame interpolation methods as illustrated in Figure 1. In our proposed setup, we place a pair of RGB cameras on the top corner of the monitor to capture a left and right view of the video participant. The baseline of the cameras is flexible and can be determined based on the size of the monitor. Then, a facial landmark detection algorithm is applied to the captured image for the facial landmark detection, which is used to estimate the angle of head rotation and align the image pairs. Using the detected landmarks for image alignment also makes our proposed method calibration-free. To solve the problem of distractions from the background, we propose performing background removal before video frame interpolation. We also designed a light-weight CNN-ViT hybrid video frame interpolation model that performs real-time frame interpolation of high-resolution images.

In summary, our contributions are:

- We propose a new calibration-free frontal view synthesis method that only requires two RGB cameras and a consumer graphics card, which is affordable compared to existing methods.
- Our proposed methods solve the challenges of sophisticated

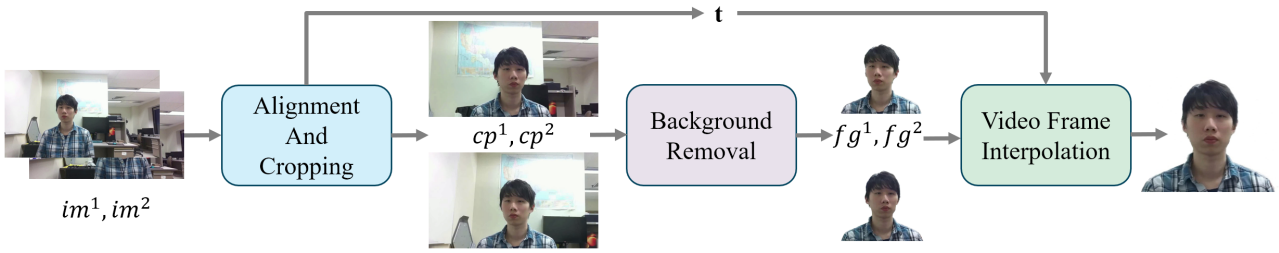


Figure 2: The overall pipeline of our proposed frontal view synthesis method. Our pipeline consists of three major components: alignment and cropping, background removal, and video frame interpolation.

calibration and distraction from the background by introducing facial landmark detection and background segmentation into the pipeline.

- We propose a novel CNN-ViT hybrid model architecture for video frame interpolation, capable of performing high-quality video frame interpolation in real-time.

Related Works

Frontal View Synthesis aims to generate a forward-facing view of a subject from non-frontal or arbitrary angles, playing a crucial role in applications such as face recognition, virtual avatars, and photorealistic reenactment. Traditional approaches rely on 3D Morphable Models (3DMMs) [6], [7], which fit a parametric face model to the input and re-render it from a frontal perspective. However, these methods often struggle with the preservation of texture and identity consistency [8]. Deep learning-based techniques have significantly improved frontalization quality, with Generative Adversarial Networks (GANs) [9]–[11] learning to synthesize realistic frontal views while maintaining identity features. More recent neural rendering methods, such as NeRF-based and implicit function [12], [13] approaches, further enhance synthesis by leveraging multi-view consistency and view-dependent effects. Additionally, self-supervised and few-shot learning strategies [14] enable robust synthesis even under extreme poses and occlusions. Despite these advancements, challenges remain in handling occlusions, extreme lighting variations, and preserving fine details, motivating research into hybrid techniques that combine explicit geometry priors with neural synthesis.

Facial Landmark Detection is a fundamental task in computer vision, aiming to localize key facial points for applications in face recognition, expression analysis, and 3D face reconstruction. The advent of deep learning significantly improved landmark detection, with Convolutional Neural Networks (CNNs) [15] and Heatmap-based Regression [16] achieving robust and accurate predictions. More recent works leverage Graph Neural Networks (GNNs) [17] and Vision Transformers (ViTs) [18] to model spatial relationships between landmarks and enhance generalization across diverse datasets. Additionally, self-supervised and multi-task learning approaches [19] enable landmark detection under extreme conditions, such as profile views and occlusions.

Portrait Segmentation is a key task in computer vision, aiming to separate a subject from the background for applications such as virtual backgrounds, augmented reality (AR), and computational photography [20]. Traditional approaches relied on color-based segmentation and graph-cut methods, which struggled with complex backgrounds and varying lighting conditions. The advent of deep learning led to significant improvements, with Fully Convolutional Networks (FCNs) and U-Net architectures [21] enabling accurate pixel-wise segmentation. More recent approaches leverage wide receptive field and self-attention mechanisms [22], [23], improving boundary refinement and robustness to occlusions. Additionally, lightweight models [24], [25] optimized for mobile devices enable real-time portrait segmentation for AR applications.

Video Frame Interpolation (VFI) aims to synthesize intermediate frames between existing ones, enhancing temporal resolution for applications such as slow motion generation, frame rate upscaling, and novel view synthesis. Early methods relied on optical flow estimation [26] to model motion between consecutive frames, followed by motion compensation and warping. However, forward warping optical flow-based methods often struggled with occlusions and complex motion [27]. Deep learning significantly advanced VFI, with learnable backward warping flow prediction and deep recurrent architectures [28] learning motion priors for more accurate interpolation. Recent methods, such as Transformer-based [29], further improve performance by considering long-range dependencies and scene geometry. Despite these advancements, challenges remain in handling large motions, especially in the areas where there is no correspondence in the input image pairs [4].

Method

Overall Pipeline

We propose a frontal view synthesis pipeline composed of three key modules: Alignment and Head Pose Estimation, Background Removal, and Video Frame Interpolation, as illustrated in Figure 2. When input images, denoted as im^1, im^2 are captured using two cameras positioned at the left and right corners of the screen, facial landmark detection is first applied to both images. The detected landmarks facilitate aligned cropping and head pose estimation. The relative spatial position of the novel view denoted as $t \in (0,1)$, is determined based on the yaw angle of the head poses. To eliminate background distractions, a portrait segmentation model is applied to the cropped image pairs. Finally, the cleaned image pairs are processed by the video frame interpolation module to generate the synthesized novel view. Since our model is designed for real-time applications, each component is carefully selected to balance both accuracy and efficiency.

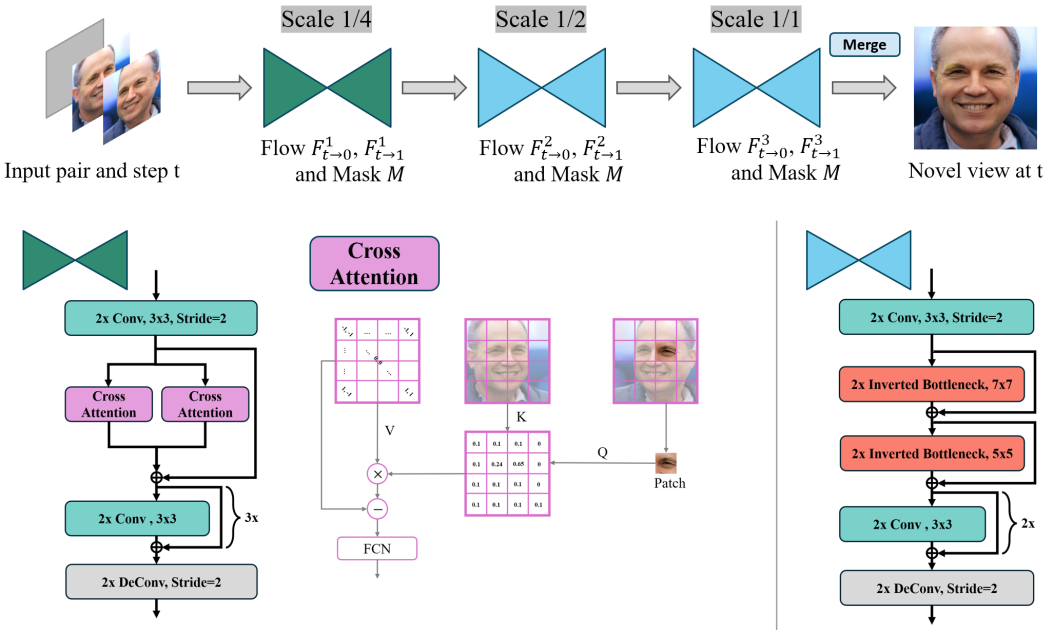


Figure 3: The architecture of our video frame interpolation model.

Alignment

The alignment and head pose estimation module is constructed around a facial landmark detection model PFLD [30], which is accurate and efficient. The input of the PFLD is an RGB image of size 112×112 and we extract five points: the left corner of the left eye, the right corner of the right eye, the nose, and left and right mouth corners, from the predicted landmarks. The pixel coordinates of the points are denoted as $(x_n, y_n), n \in \{1..5\}$. To support image pairs taken from webcams with different focal lengths, we first resize the image with a smaller focal length. The scaling factor f_s is calculated using:

$$f_s = \frac{y_1^1 + y_2^1 - y_4^1 - y_5^1}{y_1^2 + y_2^2 - y_4^2 - y_5^2} \tag{1}$$

We adapt the ROI crop algorithm [25] to work with our detected nose points $\{x_3, y_3\}$. We calculate the size, $\{x', y'\}$ of the bounding box using:

$$\begin{aligned} up &= min(y_3^1, y_3^2) \\ bottom &= H - max(y_3^1, y_3^2) \\ left &= min(x_3^1, x_3^2) \\ right &= W - max(x_3^1, x_3^2) \\ x' &= min(left, right) \times 2 \\ y' &= up + bottom \end{aligned} \tag{2}$$

where H and W denote the height and width of the target image after cropping. The left and right images after alignment are denoted as $\{cp^1, cp^2\}$.

Relative Camera Pose Estimation

To interpolate the left and right camera view of the video participant to a frontal camera view, it is essential to estimate the camera poses. Since we have the facial landmark points, we can fit the landmarks from the left and right images to a pre-defined 3D head to solve for the poses. We apply the algorithm described in [31] to the left and right images $\{cp^1, cp^2\}$ to estimate the yaw angles of the left and right camera with respect to the video participant, denoted as $\{yaw^1, yaw^2\}$, in degrees. Finally, we estimate the target camera position t at $yaw^t = 0$ by normalizing $\{yaw^1, yaw^2\}$ to zero to one:

$$t = \frac{-yaw^1}{yaw^2 - yaw^1} \tag{3}$$

Background Removal

Video frame interpolation models work by finding the feature correspondence in the input image pairs and synthesis the intermediate novel image. This process, however, is prone to error where foreground objects occlude backgrounds [27]. Previous works have found that such artifacts greatly undermine the visual fidelity of human portraits by ruining the clarity and completeness of face boundaries [4]. To mitigate this problem, we propose to perform background removal prior to performing view interpolation. We incorporated the portrait segmentation and matting model GRIB [23] into our frontal view synthesis pipeline due to its lightweight and robustness. The background removal module takes $\{cp^1, cp^2\}$ as input and outputs the foreground pairs $\{fg^1, fg^2\}$.

Video Frame Interpolation

To achieve real-time frame interpolation at HD (1280×720) and FHD (1920×1080) resolutions, we design a lightweight optical flow base frame interpolation model illustrated in Figure 3. Our model features a 3-block iterative design (base block, refinement block, refinement block) to estimate the optical flow from coarse to fine on a scale of 1/4, 1/2, and full input resolution. The input to the model is the channel-wise concatenation of image pair $\{fg^1, fg^2\}$ and step t. In each block, our model predicts a pair of backward flows from the novel view at t to $\{fg^1, fg^2\}$ and a con-

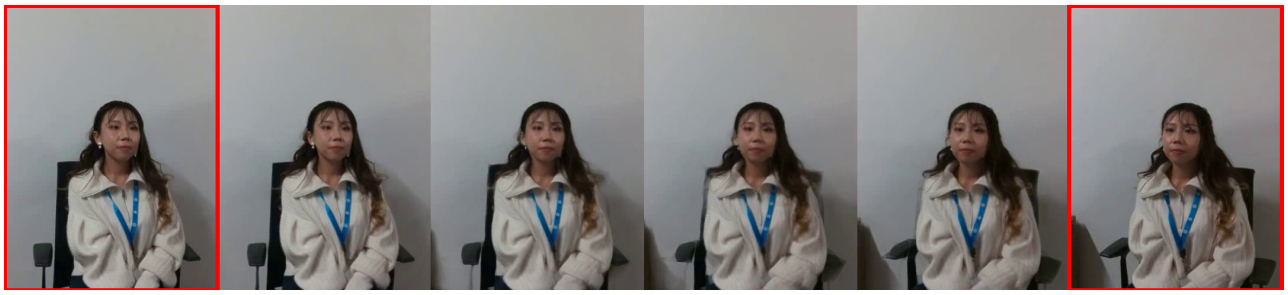


Figure 4: A visual example of the interpolation result from our VFI model. The left and right most images in the red bounding box are the input image pairs. The three intermediate frames in the middle are interpolated by our model.

fidence mask M which is the predicted probability of $\{fg^1, fg^2\}$'s pixel contribution to the synthesized result. The final output image is produced by merging the warped left and right images using the predicted mask.

Base Block has a CNN-ViT hybrid architecture, which consists of two stride convolution layers for downsampling, two cross-attention motion extractors, six convolution layers for backward flow prediction, and two deconvolution layers for upsampling. We choose to use a kernel size of 3×3 for all the convolution layers following [28] for efficient feature extraction. After the downsampling layers, the features are separated into $\{fe^1, fe^2\}$ evenly by channel. Each feature patch in the query (Q) is attended with all keys (K) extracted from the other feature, denoted as QK. The value (V) is constructed from an evenly spaced mesh grid ranging from 0 to 1 on the columns and rows subtracting the relative position of each patch in Q. The matrix multiplication of QK and V results in the relative motion map. The two motion maps are concatenated channel-wise before feeding into the consecutive convolution layers. We divide the base block into groups of two and incorporate skip connections to preserve information from previous layers around the feature extractor and flow prediction groups.

Refinement Block operates on higher spatial resolutions, thus, we select a full convolution architecture for efficiency. Similar to the base block, each refinement block contains a group of downsampling layers, two groups of inverted bottleneck layers for motion extraction, two groups of convolution layers for flow prediction, and a group of upsampling layers. To extract motion from wider ranges, we set the kernel size of the inverted bottleneck groups to 7×7 and 5×5.

Experiments

Dataset

We use 4 datasets in total for training and evaluation.

1) Vimeo90K [32] consists of two subsets, Triplet and Septuplet, and contains video sequences at a fixed resolution of 448×256. The Triplet subset includes 51,313 sequences with three consecutive video frames, while the 64,612 video sequences in the Septuplet subset each contain seven frames. We crop the image pairs to 256×256 during training.

2) X4K1000FPS [33] consists of 2160p video samples captured at high frame rate. The original dataset contains 4,408 training clips of resolution 768×768 and 15 testing clips of resolution 2048×1024. We crop the image pairs to 512×512 during training.

3) ZJU-MoCap [34] includes 8,843 clips of human body movement videos at a fixed resolution of 512×512 when divided into lengths of 20 frames.

4) EG3D-syn contains 75,000 clips of portrait images generated at 9 different camera positions. We follow [4] to generate all the clips using EG3D [35].

Training

We train our video frame interpolation method via a two-stage method. In the first stage, we train the model on the dataset Vimeo90K and X4K1000FPS, where a wide variety of scenes appear. We train the model for 200 epochs interleaved on 4 GPUs with a total batch size of 64. The model is trained using a cosine learning rate scheduler with an initial rate of 4×10^{-4}. In the second stage, we finetune the model on ZJU-MoCap and EG3D-syn dataset for 10 epochs interleaved with a fixed learning rate of 4×10^{-5}.

Evaluation

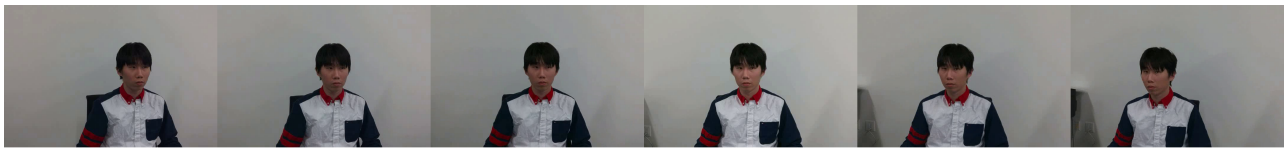
We evaluate our VFI model against other recent real-time frame interpolation models including RIFE [28], M2M [36], and IFRNet[37] on Vimeo90K Triplet and 4K1000FPS. Figure 4 shows an example of video interpolation results generated by our VFI model.

Fixed time step interpolation refers to synthesizing the middle frame between two input images. We evaluate the accuracy of VFI methods' fixed time step interpolation on the Vimeo90K Triplet dataset. We follow the testing procedure described in [28] to perform a quantitative evaluation of the interpolation results. To achieve a fair comparison and demonstrate our model's capability, the evaluation is performed after the first training stage. We report the Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and frames per second (FPS) evaluated on a Nvidia RTX 3090. As shown in Table 1, our method achieves the highest PSNR and SSIM. Our model also achieves the second-highest FPS, indicating our model is efficient and fast on low-resolution image pairs.

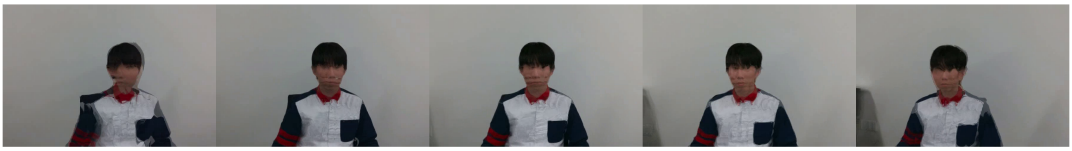
Methods	PSNR	SSIM	FPS
RIFE [28]	35.61	0.9779	**379.7**
M2M [36]	35.47	0.9778	223.1
IFRNet [37]	<u>35.80</u>	<u>0.9794</u>	339.6
Ours	**35.94**	**0.9795**	<u>365.1</u>

Table 1: The fix time step interpolation evaluation on Vimeo90K Triplet dataset. The best and second-best results are marked in **bold** and <u>underline</u>.

Arbitrary time step interpolation refers to synthesizing in-



(a) Input frames



(b) Interpoalted frames without alignment

(c) Interpolated frames after alignment and cropping

Figure 5: Examples of visual comparison to show the effectiveness of alignment and cropping. Please zoom in for a better viewing experience

termediate frames at arbitrary steps between the two input images. We evaluate the accuracy of VFI models's arbitrary time step interpolation on the X4K1000FPS dataset. We follow the testing procedure proposed by [33] to assess the interpolation accuracy at every eighth interval in each testing clip. We perform the interpolation results on the resolution of 2K and 4K and report the PSNR and FPS of both resolutions. As shown in Table 2, our model achieves the second-best performance in terms of both accuracy and efficiency.

Methods	PSNR-2K	FPS-2K	PSNR-4K	FPS-4K
RIFE	31.43	**22.5**	30.58	**5.49**
M2M	**32.13**	15.8	**30.88**	4.33
IFRNet	31.53	19.2	30.46	4.27
Ours	<u>31.88</u>	<u>19.9</u>	<u>30.80</u>	<u>4.79</u>

Table 2: The arbitrary time step interpolation evaluation on X4KF1000FPS dataset. The best and second-best results are marked in **bold** and <u>underline</u>.

Ablation Study

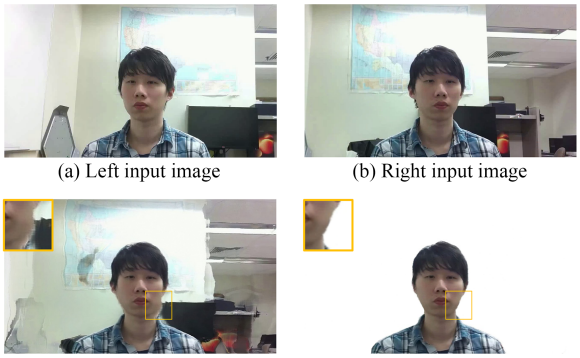
In this section, we perform ablation studies on the other two major components of our proposed method: The alignment module and the Background removal module.

Alignment Module

In this experiment, we interpolate a six-frame test video to illustrate the contribution of the alignment module. To eliminate the possible distortion caused by complex background geometries, we record this video in front of a clean white background. Figure 5 shows that the interpolation model fails to produce good-quality results due to the large motion in the original frames. However, with alignment and cropping, the model is capable of producing high-quality interpolated results.

Background Removal Module

In this experiment, we use the IFNet [28] model as a baseline to demonstrate the effectiveness of the background removal module. We record a pair of input videos of a person in front of a complex background and perform frame interpolation using the baseline model on the aligned input images. Figure 6 shows the case where the edge of the participant's face in the prediction is distorted due to the distraction from background objects. However, when we apply background removal to the input image pairs before interpolation, we are able to produce better results.



(a) Left input image (b) Right input image

(c) Baseline result w/o backgorund removal (d) Baseline result w/ backgorund removal

Figure 6: Example of the effectiveness of the background removal module when the frame interpolation model fails to predict good-quality results.

Conclusion

In this paper, we propose a frontal view synthesis method for immersive video conferencing using dual-camera capture and frame interpolation. Our method eliminates the drawbacks of being 1) difficult to calibrate, 2) prone to complex backgrounds, and 3) computationally expensive to run, as found in existing frontal view synthesis methods. We also architecture a lightweight video frame interpolation model capable of generating high quality novel views. Our proposed method exhibits the potential to be further expanded to other applications including 3D telepresence and virtual reality.

References

[1] J. Lawrence, R. Overbeck, T. Prives, T. Fortes, N. Roth, and B. Newman, "Project starline: A high-fidelity telepresence system," in *ACM SIGGRAPH 2024 Emerging Technologies*, 2024, pp. 1–2.

[2] A. Trevithick, M. Chan, M. Stengel, *et al.*, "Real-time radiance fields for single-image portrait view synthesis," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–15, 2023.

[3] Z. Ye, T. Zhong, Y. Ren, *et al.*, "Real3d-portrait: One-shot realistic 3d talking portrait synthesis," *arXiv preprint arXiv:2401.08503*, 2024.

[4] W. Xu, Y. Shen, Q. Lin, J. P. Allebach, and F. Zhu, "Pose guided portrait view interpolation from dual cameras with a long baseline," in *2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2024, pp. 1–6.

[5] Z. Liu, W. Jia, M. Yang, P. Luo, Y. Guo, and M. Tan, "Deep view synthesis via self-consistent generative network," *IEEE Transactions on Multimedia*, vol. 24, pp. 451–465, 2021.

[6] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, "A 3d morphable model learnt from 10,000 faces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5543–5552.

[7] L. Tran and X. Liu, "Nonlinear 3d face morphable model," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7346–7355.

[8] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou, "3d face morphable models" in-the-wild"," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 48–57.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[10] A. Kammoun, R. Slama, H. Tabia, T. Ouni, and M. Abid, "Generative adversarial networks for face generation: A survey," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–37, 2022.

[11] C. Rong, X. Zhang, and Y. Lin, "Feature-improving generative adversarial network for face frontalization," *IEEE access*, vol. 8, pp. 68 842–68 851, 2020.

[12] Y. Hong, B. Peng, H. Xiao, L. Liu, and J. Zhang, "Headnerf: A real-time nerf-based parametric head model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 374–20 384.

[13] J. Sun, X. Wang, Y. Zhang, *et al.*, "Fenerf: Face editing in neural radiance fields," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7672–7682.

[14] J. Zhang, X. Li, Z. Wan, C. Wang, and J. Liao, "Fdnerf: Few-shot dynamic neural radiance fields for face reconstruction and expression editing," in *SIGGRAPH Asia 2022 Conference Papers*, 2022, pp. 1–9.

[15] Y. Wu, T. Hassner, K. Kim, G. Medioni, and P. Natarajan, "Facial landmark detection with tweaked convolutional neural networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 3067–3074, 2017.

[16] A. Bulat, E. Sanchez, and G. Tzimiropoulos, "Subpixel heatmap regression for facial landmark localization," *arXiv preprint arXiv:2111.02360*, 2021.

[17] Q. T. Ngoc, S. Lee, and B. C. Song, "Facial landmark-based emotion recognition via directed graph neural network," *Electronics*, vol. 9, no. 5, p. 764, 2020.

[18] H. Li, Z. Guo, S.-M. Rhee, S. Han, and J.-J. Han, "Towards accurate facial landmark detection via cascaded transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4176–4185.

[19] Z. Sun, C. Feng, I. Patras, and G. Tzimiropoulos, "Lafs: Landmark-based facial self-supervised learning for face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1639–1649.

[20] W. Xu, Y. Shen, Q. Lin, J. P. Allebach, and F. Zhu, "Efficient real-time portrait video segmentation with temporal guidance," *Electronic Imaging*, vol. 34, pp. 1–7, 2022.

[21] S.-H. Zhang, X. Dong, H. Li, R. Li, and Y.-L. Yang, "Portraitnet: Real-time portrait segmentation network for mobile device," *Computers & Graphics*, vol. 80, pp. 104–113, 2019.

[22] Y. Shen, W. Xu, Q. Lin, J. P. Allebach, and F. Zhu, "Real-time end-to-end portrait and in-hand object segmentation with background fusion," in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, IEEE, 2023, pp. 242–247.

[23] Y. Shen, W. Xu, Q. Lin, J. P. Allebach, and F. Zhu, "Grib: Combining global reception and inductive bias for human segmentation and matting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5576–5585.

[24] Y. Shen, W. Xu, Q. Lin, J. P. Allebach, and F. Zhu, "Depth assisted portrait video background blurring," *Electronic Imaging*, vol. 35, pp. 1–6, 2023.

[25] W. Xu, Y. Shen, Q. Lin, J. P. Allebach, and F. Zhu, "Exploiting temporal information in real-time portrait video segmentation," in *Proceedings of the 4th International Workshop on Human-Centric Multimedia Analysis*, 2023, pp. 33–39.

[26] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang, "Depth-aware video frame interpolation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3703–3712.

[27] S. Niklaus and F. Liu, "Softmax splatting for video frame interpolation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5437–5446.

[28] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou, "Real-time intermediate flow estimation for video frame interpolation," in *European Conference on Computer Vision*, Springer, 2022, pp. 624–642.

[29] G. Zhang, Y. Zhu, H. Wang, Y. Chen, G. Wu, and L. Wang, "Extracting motion and appearance via inter-frame attention for efficient video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5682–5692.

[30] X. Guo, S. Li, J. Yu, *et al.*, "Pfld: A practical facial landmark detector," 2019.

[31] S. Mallick, *Head pose estimation using opencv and dlib*, 2016. [Online]. Available: https://learnopencv.com/head-pose-estimation-using-opencv-and-dlib/.

[32] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, pp. 1106–1125, 2019.

[33] H. Sim, J. Oh, and M. Kim, "Xvfi: Extreme video frame interpolation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14 489–14 498.

[34] Q. Shuai, C. Geng, Q. Fang, *et al.*, "Novel view synthesis of human interactions from sparse multi-view videos," in *ACM SIGGRAPH 2022 conference proceedings*, 2022, pp. 1–10.

[35] E. R. Chan, C. Z. Lin, M. A. Chan, *et al.*, "Efficient geometry-aware 3d generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 123–16 133.

[36] P. Hu, S. Niklaus, S. Sclaroff, and K. Saenko, "Many-to-many splatting for efficient video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3553–3562.

[37] L. Kong, B. Jiang, D. Luo, *et al.*, "Ifrnet: Intermediate feature refine network for efficient frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1969–1978.

Yezhi Shen is a Ph.D. student of Electrical and Computer Engineering at Purdue University, West Lafayette, Indiana. He received a B.S. degree in Electrical and Computer Engineering from Purdue University in 2021. His main research area covers dense prediction, perception, and 3D reconstruction in computer vision.

Md Adnan Faisal Hossain is pursuing a Ph.D. degree in Electrical and Computer Engineering at Purdue University, West Lafayette, Indiana. He received his B.S. degree in Electrical and Electronics Engineering from Bangladesh University of Engineering and Technology in 2022. He works as a graduate research assistant in the Video and Image Processing Laboratory at Purdue University. His main research area covers deep learning model compression, learned image compression, and generative image compression.

Weichen Xu received his Ph.D. degree in Electrical and Computer Engineering at Purdue University. He works as a Research Engineer at HP Inc. His research interests include video processing and deep learning.

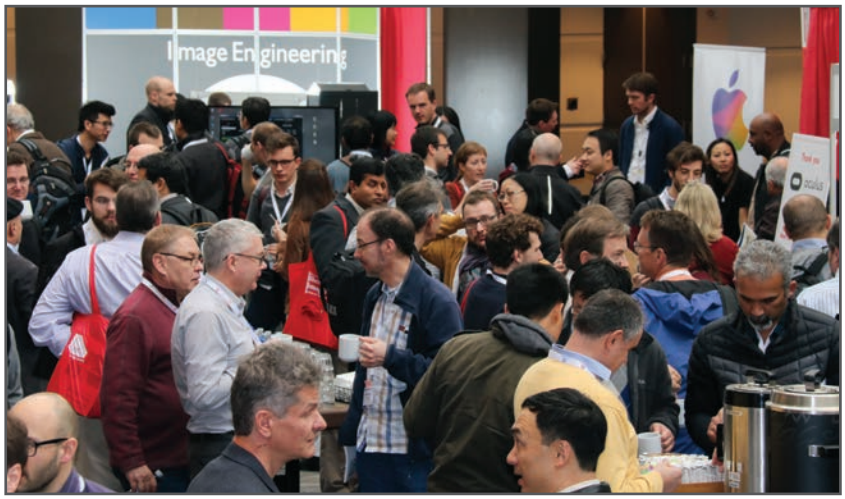
Qian Lin is an HP Fellow working on computer vision and deep learning research. She is also an adjunct professor at Purdue University. She joined Hewlett-Packard Company in 1992. She received her BS from Xi'an Jiaotong University in China, MS from Purdue University, and Ph.D. in Electrical Engineering from Stanford University. She is an inventor/co-inventor for over 45 issued patents. She was awarded a fellowship from the Society of Imaging Science and Technology (IS&T) in 2012, Outstanding Electrical Engineer by the School of Electrical and Computer Engineering of Purdue University in 2013, and the Society of Women Engineers Achievement Award in 2021.

Fengqing Zhu is an Associate Professor of Electrical and Computer Engineering at Purdue University, West Lafayette, Indiana. She received her B.S. (with highest distinction), M.S., and Ph.D. degrees in Electrical and Computer Engineering from Purdue University. She is the recipient of an NSF CISE Research Initiation Initiative (CRII) award in 2017, a Google Faculty Research Award in 2019, and an ESI and trainee poster award for the NIH Precision Nutrition workshop in 2021. She is a senior member of the IEEE.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org