# SVFR: A Unified Framework for Generalized Video Face Restoration

## Supplementary Material

In the supplementary material, Sec.A first presents more results of our pilot study. Then, Sec.B shows details about our implementations in our unified face restoration framework training. Following, Sec.C provides more visual results of comparisions among SOTA methods, and Sec.D presents the inference speed and memory consumption. Finally, Sec. E presents more related works.

## A. Face Restoration Pilot Study

In the main artical, we discuss a pilot experiment designed to validate the mutual benefits among different facial restoration tasks. Specifically, our task pool includes BFR, colorization, and inpainting. To examine whether prior knowledge from one task can benefit another, we use GPEN as the baseline method and train under two distinct settings: (1) training from the weights of pretrained StyleGANv2, and (2) transfer learning, where the model is pretrained on one task and fine-tuned on the target task. The results demonstrate that models trained without restoration prior task-specific knowledge exhibit weaker performance. In contrast, models trained with restoration prior knowledge achieve better results, highlighting that knowledge sharing among tasks positively contributes to the performance of individual facial restoration subtasks.

As shown in Fig.1, BFR trained with transfer learning demonstrates more specific structures compared to training initialized with pretrained StyleGANv2. This indicates that pretraining on the inpainting task benefits BFR training by enhancing the model's ability to restore face. For example, results with transfer learning demonstrate a more complete and realistic tooth structure, while the counterpart without transfer learning exhibits noticeable distortions in the tooth region. Similarly, the colorization task, when trained with BFR pretraining, exhibits improved performance, accurately coloring the appropriate regions without overspreading. For example, without transfer learning, the results mistakenly color certain parts of the hair with unnatural hues, whereas with transfer learning, this issue is not occur. Moreover, inpainting trained with BFR pretraining becomes more precise, focusing specifically on the masked regions. Without transfer learning, the beard details are not accurately restored, and noticeable areas of incomplete restoration appear.

This indicates that leveraging the prior knowledge from related tasks not only accelerates the learning process but also enhances the performance of the target task, providing strong evidence for the effectiveness of multi-task learning in video face restoration.

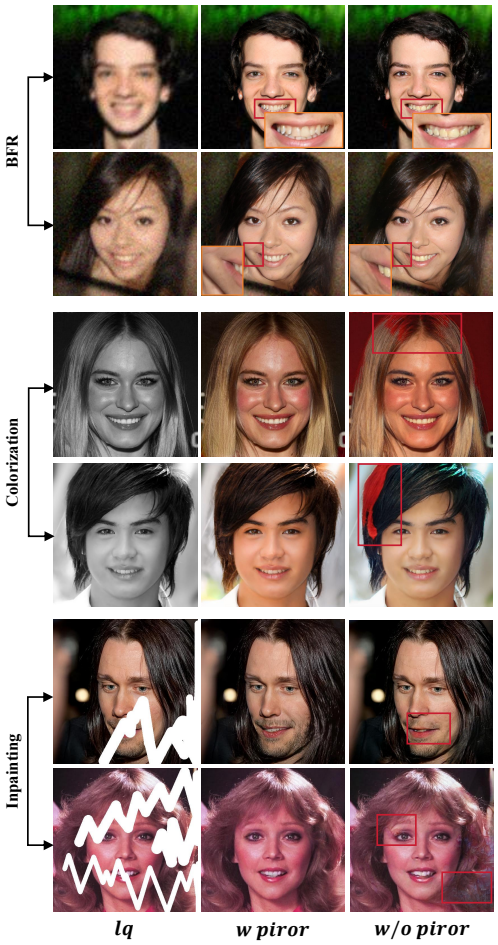


Figure 1. **The results of our pilot study.** For BFR (top two rows), transfer learning enhances realism in features like teeth and eyes. In colorization (middle two rows), it improves sensitivity to human regions, avoiding errors in areas like hair. For inpainting (bottom two rows), transfer learning restores fine facial textures, while models without it leave noticeable artifacts.

## B. Implementation Details

### B.1. Training Network Details

As discussed in the main text, our model incorporates both a multi-task training framework and a facial structure prior framework. Below, we provide a detailed description of the corresponding model architecture.

**Chosen layer.** For Unified Latent Regularization (ULR), we flatten mid-block features in the height and width dimensions to compute loss, leveraging their smaller spatial size to reduce computational load. Besides, both ULR and facial prior learning (FPL) require relatively deeper features

Table 1. Ablation study on $\lambda_1$ (for $\mathcal{L}_{ULR}$) and $\lambda_2$ (for $\mathcal{L}_{piror}$) hyperparameters.

| Methods | BFR / Colorization / Inpainting | | | | | |
|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | IDS↑ | VIDD↓ | FVD↓ |
| $\lambda_1 = 0, \lambda_2 = 0$ | 28.936 / 22.921 / 28.303 | 0.854 / 0.870 / 0.898 | 0.242 / 0.274 / 0.156 | 0.881 / 0.979 / 0.875 | 0.489 / 0.501 / 0.511 | 98.781 / 223.168 / 101.146 |
| $\lambda_1 = 0.1, \lambda_2 = 0$ | 28.695 / 22.788 / 28.004 | 0.849 / 0.861 / 0.896 | 0.245 / 0.277 / 0.161 | 0.877 / 0.976 / 0.875 | 0.504 / 0.503 / 0.515 | 102.146 / 229.516 / 104.102 |
| $\lambda_1 = 0.01, \lambda_2 = 0$ | 29.296 / 22.987 / 28.337 | 0.859 / 0.886 / 0.900 | 0.225 / **0.270** / 0.155 | 0.884 / 0.978 / 0.879 | 0.486 / 0.498 / 0.508 | 90.353 / 214.846 / 93.615 |
| $\lambda_1 = 0.01, \lambda_2 = 0.5$ | 29.077 / 22.908 / 28.164 | 0.851 / 0.878 / 0.897 | 0.233 / 0.273 / 0.158 | 0.880 / 0.977 / 0.876 | 0.491 / 0.498 / 0.511 | 94.172 / 220.583 / 98.147 |
| $\lambda_1 = 0.01, \lambda_2 = 0.1$ | **29.563 / 23.079 / 29.119** | **0.862 / 0.896 / 0.904** | **0.223** / 0.272 / **0.153** | **0.902 / 0.980 / 0.888** | **0.479 / 0.497 / 0.504** | **89.316 / 204.260 / 88.354** |

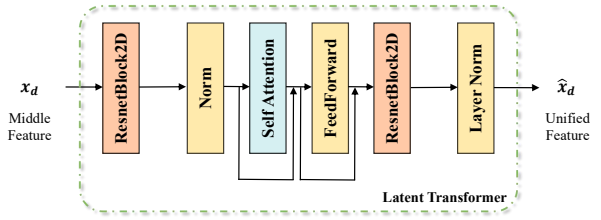


Figure 2. The structure of our Latent Transformer module.

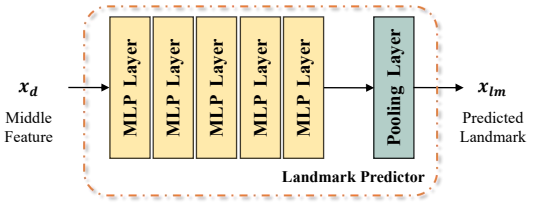


Figure 3. The structure of our Landmark Predictor module.

to capture robust intermediate information, while avoiding layers too close to the output to prevent compromising visual quality. Mid-block features offer the best balance.

**Latent Transformer module.** In multitask training, multiple tasks share a common feature space, requiring the model not only to optimize for individual tasks but also to maintain consistency within the shared feature space. To address this, we designed the Latent Transformer module, as show in Fig.2, which maps intermediate features $x_d$ from the U-Net to a unified latent space. By calculating contrastive loss on these unified features, the module effectively learns shared representations across different tasks, improving the model's ability to generalize and enhance performance.

**Landmark Predictor module.** To enhance the structural consistency of facial restoration results, we introduced a Landmark Predictor module. This module takes intermediate features from the U-Net and predicts 68 facial landmarks through the predictor. The detailed structure is illustrated in Fig.3.

### B.2. Ablation Study

As mentioned in the main text, our objective function is:

$$\mathcal{L} = \mathcal{L}_{noise} + \lambda_1 \mathcal{L}_{ULR} + \lambda_2 \mathcal{L}_{prior}. \quad (1)$$

Next, we empirically set the hyperparameters for training and conduct ablation experiments on the $\lambda_1$ and $\lambda_2$ hyperparameters. The results, shown in Tab.1, demonstrate that $\lambda_1 = 0.01$ and $\lambda_2 = 0.1$ yield the best performance.

### B.3. Data Filtering

To enhance the quality of our training data, we filtered training datasets (VoxCeleb2 [5], CelebV-Text [26], and VFHQ [23]) rigorously. We first extracted square face bounding boxes and scaled them by a factor of 0.2. Then we crop and filter out those with resolutions below 512. Subsequently, we applied the image quality assessment method ARNIQA [1], using a model trained on "live" data, to further select frames with scores above 0.75. This process yielded a high-quality video dataset comprising 20,000 clips.

## C. More Results

We conducted a comprehensive comparison with other methods on the VFHQ-test public dataset across three tasks: BFR, Colorization, and Inpainting. Additionally, we collected real-world data to further validate the effectiveness of our approach. The videos mentioned above are included in the project website https://wangzhiyaoo.github.io/SVFR/.

## D. Inference Speed

Diffusion models often suffer from high inference time and memory consumption. To evaluate the efficiency of our method, we compared its speed and GPU memory usage against other approaches. Experiments were conducted on an RTX 3090 GPU, generating 100 video frames (see Tab.2).

Table 2. Inference speed and memory.

| Methods | GPEN | CodeFormer | PGDiff | KEEP | PGTFormer | Ours |
|---|---|---|---|---|---|---|
| Time(s) | 47.39 | 47.58 | 279.08 | 15.17 | 34.58 | 250.64 |
| Memory-Usage(MiB) | 1646 | 1604 | 4884 | 15030 | 4062 | 16824 |

## E. Related Works

### E.1. Video Colorization

Video colorization involves adding color to grayscale frames while maintaining temporal coherence. Existing methods fall into three categories: post-processing techniques, exemplar-guided colorization, and reference-based approaches. Post-processing methods use temporal filters to reduce flickering but often result in desaturated colors [3, 12]. Exemplar-guided methods propagate user-provided scribbles or transfer colors from reference images, relying

on optical flow, which can introduce artifacts due to flow inaccuracies [10, 14, 19, 25]. Reference-based methods use a single colored frame to colorize subsequent frames, leveraging either hand-crafted features or deep learning for temporal correspondence [2, 9, 10, 17, 21, 22, 28]. In comparison with these works, we in this paper propose a unified framework for all three visual tasks with simple network structure and improved performance.

### E.2. Video Inpainting

Advances in video inpainting have largely been driven by methods that fill in masked regions by borrowing content from unmasked regions in other frames, known as content propagation methods. These methods typically use optical flow estimates [6, 8, 11, 24], self-attention [13, 16, 18, 27], or a combination of both [15, 29, 30] to propagate pixel values or learned features across frames. While these methods often produce visually compelling results, especially in tasks where the masked region is visible in nearby frames, they struggle with heavy camera motion, large masks, or tasks requiring semantic understanding of the video content. More recent work has utilized diffusion models for video inpainting. Gu et. al. [7] combines a video diffusion model with optical flow guidance, following a similar content propagation approach. Chang et. al. [4] uses a latent diffusion model [20] to remove the agent's view of itself from egocentric videos for robotics applications. This is framed as an image inpainting task, where the goal is to remove the agent from a single video frame conditioned on previous frames, resulting in a lack of temporal consistency.

## References


[1] Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. Arniqa: Learning distortion manifold for image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. 2

[2] Nir Ben-Zrihem and Lihi Zelnik-Manor. Approximate nearest neighbor fields in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5233–5242, 2015. 3

[3] Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. Blind video temporal consistency. *ACM Transactions on Graphics (TOG)*, 34(6):1–9, 2015. 2

[4] Matthew Chang, Aditya Prakash, and Saurabh Gupta. Look ma, no hands! agent-environment factorization of egocentric videos. In *NeurIPS*, 2023. 3

[5] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 2

[6] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 3

[7] Bohai Gu, Yongsheng Yu, Heng Fan, and Libo Zhang. Flow-guided diffusion for video inpainting. *arXiv preprint arXiv:2311.15368*, 2023. 3

[8] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Temporally coherent completion of dynamic video. *ACM Transactions on Graphics*, 35(6):196, 2016. 3

[9] Vivek George Jacob and Sumana Gupta. Colorization of grayscale images and videos using a semiautomatic approach. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 1653–1656. IEEE, 2009. 3

[10] Varun Jampani, Raghudeep Gadde, and Peter V Gehler. Video propagation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 451–461, 2017. 3

[11] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5792–5801, 2019. 3

[12] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018. 2

[13] Sungho Lee, Seoung Wug Oh, DaeYeun Won, and Seon Joo Kim. Copy-and-paste networks for deep video inpainting. In *IEEE/CVF International Conference on Computer Vision*, pages 4412–4420, 2019. 3

[14] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. 2004. 3

[15] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[16] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[17] Sifei Liu, Guangyu Zhong, Shalini De Mello, Jinwei Gu, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Switchable temporal propagation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 87–102, 2018. 3

[18] Seoung Wug Oh, Sungho Lee, Joon-Young Lee, and Seon Joo Kim. Onion-peel networks for deep video completion. In *IEEE/CVF International Conference on Computer Vision*, 2019. 3

[19] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001. 3

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022. 3

[21] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by col-

orizing videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 391–408, 2018. 3

[22] Sifeng Xia, Jiaying Liu, Yuming Fang, Wenhan Yang, and Zongming Guo. Robust and automatic video colorization via multiframe reordering refinement. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 4017–4021. IEEE, 2016. 3

[23] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[24] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[25] Liron Yatziv and Guillermo Sapiro. Fast image and video colorization using chrominance blending. *IEEE TIP*, 15(5): 1120–1129, 2006. 3

[26] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[27] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *The Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3

[28] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8052–8061, 2019. 3

[29] Kaidong Zhang, Jingjing Fu, and Dong Liu. Flow-guided transformer for video inpainting. In *European Conference on Computer Vision*, pages 74–90. Springer, 2022. 3

[30] Shangchen Zhou, Chongyi Li, Kelvin C.K Chan, and Chen Change Loy. ProPainter: Improving propagation and transformer for video inpainting. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2023. 3