Universal Video Face Restoration Method Based on Vision-Language Model

Yipiao Xu XUYP@NJUST.EDU.CN
Zhenbo Song SONGZB@NJUST.EDU.CN
Jianfeng Lu LUJF@NJUST.EDU.CN
Nanjing University of Science and Technology, China

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

The video face restoration aims to restore high-quality face video from low-quality face video, but most existing methods typically focus on specific and single degradation scene such as denoising or deblurring. However, the universal video face restoration should restore face video in various degradation scenes. In this paper, we use language prompt which describes the face information including gender, appearance and expression to guide video face restoration. To enhance the applicability, we remove the language prompt by ControlNet and incorporate the human-level knowledge from vision-language models into general networks to improve the video face restoration performance and enable the universal video face restoration. In addition, we construct a degradation dataset, which contains multiple degradations in the same scene and captions which describe the face information. Our extensive experiments show that our approach achieves highly competitive performance in universal video face restoration.

Keywords: vision-language model; video face restoration; video processing; language prompt.

1. Introduction

Video face restoration is an important research in the field of computer vision Jourabloo et al. (2017); Kumar et al. (2020). It aims to restore low-quality face video in degraded scenes such as noise or blur so as to obtain high-quality face video with well vision effect and high definition Li et al. (a); Roth et al. (2016); Tzimiropoulos (2015). Unlike non-blind video face restoration with known degradation information, there is no any prior knowledge about the degradation type or parameters for blind video face restoration Wang et al. (2021). Therefore, the difficulty of blind video face restoration should be much higher Yang et al. (2020). Compared with traditional video face restoration methods, deep learning methods Zhou et al. (2022a); Li et al. (2018); Chen et al. (2018); Menon et al. (2020) can capture advanced features and semantic information better, thus achieving more accurate and effective face restoration performance. Currently, most methods only focus on specific and single degradation task such as denoising Anwar et al. (2017) or deblurring Shen et al. (2020), while the universal video face restoration should restore face video in various degradation scenes. Face information including gender, appearance and expression in degradation videos can be described in text. Learning from language prompt to improve restoration performance is an important research direction.

Figure 1: Qualitative evaluation of our method. The video face restored through our method has better data fidelity, greater continuity, better perceptual quality, higher quality facial details.

In this paper, we propose a novel approach to guide face restoration using language prompt that include gender, appearance, and expression information. To enhance its applicability, we introduce ControlNet to remove the dependency on language prompt and integrate human-level knowledge from vision-language models into general networks, improving video face restoration performance and enabling universal video face restoration. Moreover, we construct a degradation dataset that includes corresponding face caption and multiple degradations in the same scene. It demonstrates highly competitive performance in the field of universal video face restoration, as illustrated in Figure 1.

2. Background and Related Work

Face Restoration. Traditional methods for face restoration rely on the integration of prior knowledge and degraded models Chakrabarti et al. (2007); Gunturk et al. (2003); Tang and Wang (2003). The advent of deep learning techniques has brought significant advancements in face restoration through the incorporation of convolutional neural networks Huang et al.

(2017); Tuzel et al. (2016); Yu and Porikli (2016); Zhang et al. (2018a). Recent research has focused on leveraging depth priors in face image restoration, including geometric and reference priors Bulat and Tzimiropoulos (2018); Chen et al. (2021a, 2018); Dogan et al. (2019); Li et al. (2018). Furthermore, the utilization of pre-trained GANs Karras et al. (2019) as generation priors Asim et al. (2018); He et al.; Wang et al. (2021); Yang et al. (2020, 2021) has further enhanced the quality of restoration. This approach involves mapping low-quality faces into a compact, low-dimensional space defined by the pre-trained generator, effectively treating face restoration as a conditional image generation problem. Additionally, another research approach, exemplified by VQFR Gu et al., CodeFormer Zhou et al. (2022c), RestoreFormer Wang et al., and their variants Wang et al. (2023), utilizes pre-trained vector quantization (VQ) codebooks Esser et al. (2021) as learned dictionaries for facial regions, demonstrating cutting-edge achievements in blind facial restoration.

Text-to-Image generation. Text-to-image generation involves the generation of realistic images from text descriptions Duran et al. (2019), and it has garnered significant interest due to its potential applications in areas such as content creation, design, and human-computer interaction. Early work in this field focused on developing methods to translate text descriptions into corresponding visual representations using generative models. Text-to-image models such as the stable diffusion model Rombach et al. (2022) have garnered significant attention. Building upon this work, ControlNet Zhang and Agrawala enhances the diffusion network by incorporating controls to adapt to specific task conditions. InstructPix2Pix Brooks et al. (2022) proposed a method for editing images from human instructions: given aninput image and a written instruction that tells the model what to do, the model follows these instructions to edit the image. However, due to the requirement for high-precision reconstruction capability in image restoration, it cannot be directly applied to restoration tasks.

Vision-Language Models. Vision-Language Models (VLMs) have emerged as a powerful paradigm for integrating visual and textual information, enabling a wide range of applications such as image captioning, visual question answering, and cross-modal retrieval Radford et al. (2021); Chen et al. (2021b); Li et al. (b). These models aim to bridge the semantic gap between vision and language modalities, allowing for a more comprehensive understanding of multimedia content. In recent years, the advent of large-scale pre-trained language models, such as BERT Devlin et al. (2019) and GPT Brown et al. (2020), has spurred remarkable advancements in VLMs. These models demonstrated the capacity to learn rich semantic representations from text corpora Zhou et al. (2022b), leading to their adaptation for multimodal tasks through fusion with visual encodings derived from CNNs. The resulting VLMs have exhibited strong performance in tasks like image-text retrieval, where they can effectively match images with relevant textual descriptions. Moreover, efforts have been made to develop transformer-based architectures tailored specifically for vision-language tasks. For instance, ViLBERT Lu et al. (2019) and LXMERT Tan and Bansal (2019) are notable examples of VLMs designed to jointly reason about visual and textual inputs. These models employ cross-modal attention mechanisms to capture intricate interactions between visual and textual elements, enabling them to excel in tasks requiring holistic comprehension of multimodal content.

3. Method

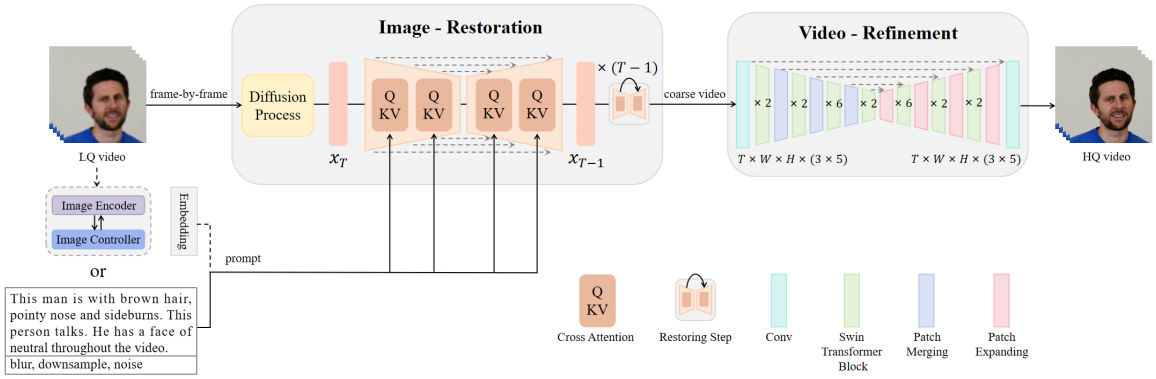

Figure 2: Overview of our method.

As illustrated in Figure 2, we input the low-quality (LQ) video along with corresponding descriptions and integrate the prompt into a general image restoration network by using cross-attention mechanism. To enhance applicability, we utilize ControlNet to remove language descriptions. Then we introduce the Video-Refinement module to further process the coarse video, which is based on Swin-Unet and can make full use of sequence information when reconstructing face.

3.1. Contrastive Controller

The image controller serves as a modified version of the CLIP image encoder, featuring additional zero-initialized connections to introduce controls to the encoder Luo et al.. The controller enables manipulation of the outputs of all encoder blocks, thereby influencing the prediction process of the image encoder. The backbone of both the encoder and the controller is ViT Dosovitskiy et al. (2020). It provides a strong foundation for capturing spatial relationships and global context within the image embeddings. To achieve discriminative and well-separated degradation-embedding spaces, we employ a contrastive objective Tian et al. (2020). This contrastive objective can learn the embedding matching process, enhance the distinction between different degradation types and facilitate more effective image restoration. By leveraging this contrastive objective, it can better align the degradation-aware representations with the underlying semantics of the images, ultimately improving the quality of image restoration.

3.2. Image Restoration

In order to integrate the prompt into a general image restoration network, we introduce a cross-attention Rombach et al. (2022) mechanism to learn semantic guidance and use IR-SDE Luo et al. (2023) as the base framework for image restoration. Its key construction includes a mean reversal stochastic differential equations, which converts high-quality images into degraded images as the mean of fixed Gaussian noise. Then, by simulating the

corresponding reverse time SDE, we can restore the origin image from low-quality images. Forward SDE for image degradation is defined as:

$$dx = \theta_t(\mu - x)dt + \sigma_t dw \tag{1}$$

where μ is the state mean, and θ_t, σ_t are time-dependent positive parameters that characterize the speed of the mean reversion and the stochastic volatility, respectively. Reverse-Time SDE for image restoration is defined as:

$$dx = [\theta_t(\mu - x) - \sigma_t^2 \bigtriangledown_x \log p_t(x)]dt + \sigma_t d\hat{\omega} \tag{2}$$

At test time, the only unknown part is the score $\bigtriangledown_x \log p_t(x)$ of the marginal distribution at time t. But during training, the ground truth, high-quality image $x(0)$ is available and thus we can train a neural network to estimate the conditional score $\bigtriangledown_x \log p_t(x \mid x(0))$.

3.3. Video Refinement

We use Swin-Unet as backbone that can make full use of sequence information when reconstructing face. Different from Cao et al. (2023), this network can work on videos. The model can enhance T frames at once exploiting spatio-temporal information. In addition, we use 3D convolution to divide the input into patches and perform pixel shuffling on the patch extension layer. The skip connection between degraded input and restored output enables the network to learn the residuals of each frame Galteri et al. (2020). It reduces the overall training time and improves its stability. The training loss is the weighted sum of pixel loss and perceptual loss defined in the VGG-19 Simonyan and Zisserman (2015) feature space. The network is trained by using 256 x 256 blocks randomly cropped from input frames. The number of frames T is fixed at 5 during training and testing.

4. Dataset Construction

Face images should exhibit various types of degradation, but existing methods only consider one type of degradation. These methods perform well on specific type of degradation problems but may fail on other types, which indicate the weak generalization ability of the models. Blind video face restoration aims to recover high-quality face images from low-quality ones without prior knowledge of the degradation type or parameters. Unlike focusing on a single degradation type, for blind video face restoration tasks, the degradation type is very complex, such as random combinations of noise, blur, low resolution, and JPEG compression artifacts et al. In addition, the blind video face restoration lacks accurate textual descriptions or clear image prompt to guide the restoration process, which hinders the model from accurately understanding the desired restoration target and increases the difficulty of the restoration process. Therefore, in this paper, we construct a video face dataset with multiple degradation types in the same scene and face captions based on CelebV-HQ Zhu et al. (2022).

4.1. Generating Degradation Video

CelebV-HQ contains 35, 666 videos, including 15, 653 identities Zhu et al. (2022). We randomly select several videos from these high-quality videos to construct video face dataset

Figure 3: Examples of degradation combination.

Table 1: The models for each single degradation type.

Degradation Types	Degradation models
blur	$y \circledast k_\sigma$
colorjitter	$(y)_{COLOR-JITTER_{b,c,s,h}}$
downsample	$(y)\downarrow_r$
jpeg	$(y)_{JPEG_q}$
noise	$y + n_\delta$

with multiple degradation types. Inspired by Wang et al. (2021), for each video, we first randomly select several types of degradation from blur, downsample, noise, jpeg, and colorjitter. Then we adopt the following degradation model to synthesize data:

$$\left\{[(y \circledast k_\sigma)\downarrow_r + n_\delta]_{JPEG_q}\right\}_{COLOR-JITTER_{b,c,s,h}} \tag{3}$$

The high quality image y is first convolved with Gaussian blur kernel k_σ followed by a downsampling operation with a scale factor r. After that, additive white Gaussian noise n_δ is added to the image. Then it is compressed by JPEG with quality factor q. Finally, it randomly jitter the brightness b, contrast c, saturation s, and hue h, in torch Tensor formats. For each video, we randomly sample σ, r, δ, q, b, c, s and h from $\{1:100\}$, $\{1.5:1.9\}$, $\{1000:3000\}$, $\{50:60\}$, $\{-20:20\}$, $\{0.5:1.5\}$, $\{0.5:1.5\}$, $\{-0.5:1.5\}$, respectively. If a certain degradation type is not selected, its corresponding parameter is zero. Table 1 summarises the models for each single degradation type. Our dataset contains 31 different degradation combinations and includes a training set of 10000 video data, a validation set of 500 video data and a testing set of 500 video data. Figure 3 shows examples of degradation combination. In addition, the degradation combination type is recorded in the caption.

4.2. Generating Face Caption

CelebV-HQ was manually labeled with 83 facial attributes, covering appearance, action, and emotion attributes for each video Zhu et al. (2022). Inspired by Yu et al. (2023), as shown in Table 2, we first classify these attributes into five kinds of category. For each video, annotated attribute information is fed into our designed template for auto-text generation. To make our template as natural as possible, we randomly select several common grammar

Table 2: The attributes of face video.

IsAttributes	blurry, young, chubby, bald
HasAttributes	pale_skin, rosy_cheeks, oval_face, receding_hairline, bangs, black_hair, blonde_hair, gray_hair, brown_hair, straight_hair, wavy_hair, long_hair, arched_eyebrows, bushy_eyebrows, bags_under_eyes, narrow_eyes, big_nose, pointy_nose, high_cheekbones, big_lips, double_chin, no_beard, sideburns, 5_o_clock_shadow, goatee, mustache, heavy_makeup
WearAttributes	eyeglasses, sunglasses, wearing_earrings, wearing_hat, wearing_lipstick, wearing_necklace, wearing_necktie, wearing_mask
ActionAttributes	blows, chews, closes_eyes, coughs, cries, drinks, eats, frowns, gazes, glares, wags_head, kisses, laughs, listens_to_music, looks_around, makes_a_face, nods, plays_instrument, reads, shakes_head, shouts, sighs, sings, sleeps, smiles, smokes, sneers, sneezes, sniffs, talks, turns, weeps, whispers, winks, yawns
EmotionAttributes	neutral, happy, sadness, anger, fear, surprise, contempt, disgust

Table 3: The synonym dictionary.

male	He, This man, The man, The person, This person
female	She, This woman, The woman, The person, This person
IsVerb	is, looks, appears to be
HaveVerb	has, is with
WearVerb	wears, is wearing
EmotionVerb	has a face of, has an expression of

structures for each attribute. In addition, in order to increase diversity of generation, texts are generated based on templates with synonym replacement. We created a synonym dictionary corresponding to face attributes, as shown in Table 3. For each video, when generating text, template randomly select corresponding word from the synonym dictionary. The examples of caption are shown in Table 4.

5. Experience

5.1. Experimental Setup

Dataset. All training and testing datasets are taken from the degradation dataset described in Section 4. We evaluate our method on the degradation dataset which contains 31 different degradation combinations. We used 50000 frames from 10000 videos for training, 2500 frames from 500 videos for evaluation, and 2500 frames from 500 videos for testing.

Evaluation Metrics. We use the Learned Perceptual Image Patch Similarity (LPIPS) Zhang et al. (2018b) and Frechet Video Distance (FVD) Unterthiner et al. (2018) as our

Table 4: The examples of caption.

The woman appears to be young, with pale skin, blonde hair, wavy hair, arched eyebrows, pointy nose and no beard. She is wearing earrings and lipstick. This person looks around, smiles and talks. The woman has an expression of happy throughout the video. : blur_noise_jpeg
The person has brown hair, arched eyebrows, pointy nose and no beard. She wears lipstick. This person wags head and talks. All the while the woman has a face of anger. : blur_downsample_noise_colorjitter
The young person has pale skin, brown hair, wavy hair, arched eyebrows, pointy nose and no beard. The woman looks around and smiles. All the while this person has an expressionof neutral. : downsample_noise_colorjitter

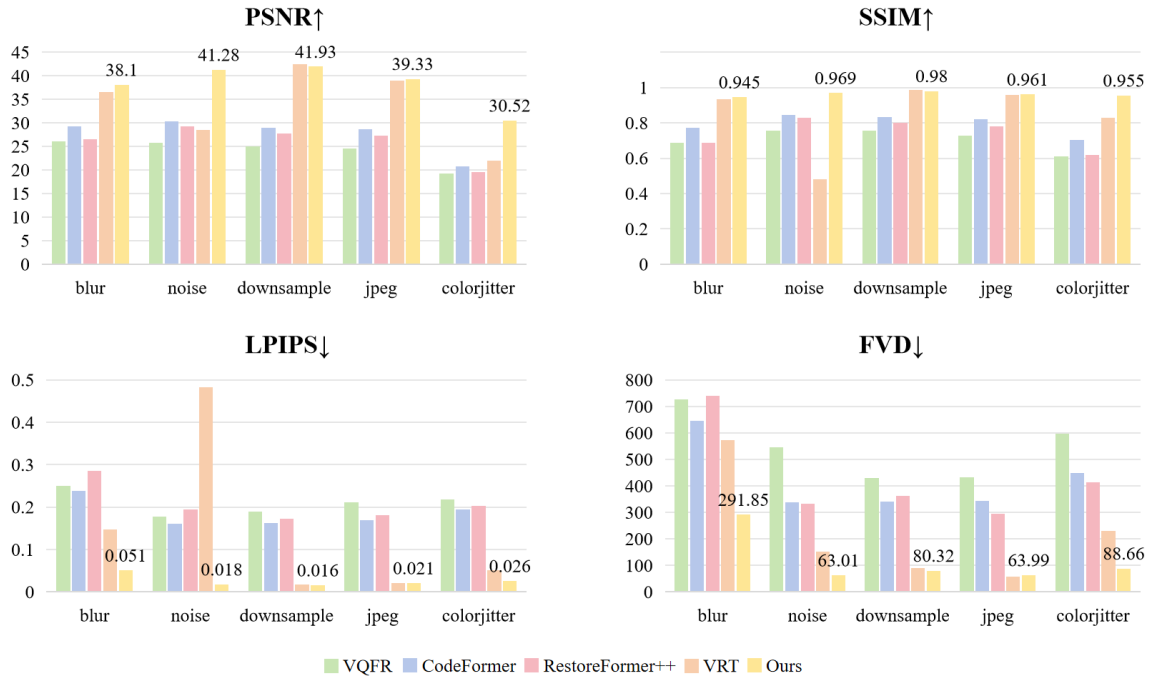

Figure 4: Comparison of five methods under a single degradation type.

metrics for perceptual evaluation and adopt two pixel-wise metrics: PSNR and SSIM Wang et al. (2004) to evaluate data fidelity of our method.

Details. During training, we resize all face images to 512×512. When training the Controller, we set $initiallearningrate = 2e-5$, $weightdecay = 0.05$. When training the image restoration model, we configured SDE's $\sigma_{max} = 50$, $T = 100$, $eps = 0.005$, and employed a cosine scheduling strategy. For training the video refinement model, we set $pixellossweight = 200$ and $perceptuallossweight = 1$.

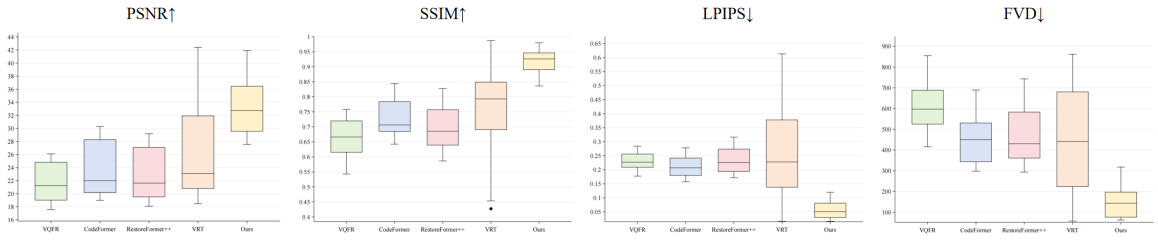

Figure 5: Comparison of five methods under 31 degradation combinations.

5.2. Comparisons with SOTA methods

Comparison approaches. We compare our method with several state-of-the-art restoration methods. VQFR Gu et al., CodeFormer Zhou et al. (2022c), and RestoreFormer++ Wang et al. (2023) are three SOTA face restoration methods that utilize pre-trained high-quality facial dictionaries as priors. The official models released by these methods were used in the experiments. We also compare a video restoration method: VRT Liang et al.. VRT is a supervised video super-resolution and deblurring deep learning method. Due to the fact that VRT cannot achieve general video restoration, we conduct experiments on video super-resolution task using their officially released model.

Degradation-specific video face restoration. The constructed degradation dataset described in Section 4 contains 31 different degradation combinations. We compare our method with other methods in these 31 degraded combinations, respectively. Figure 4 shows the comparison of five methods under a single degradation type. In addition, as shown in Figure 5, we use box plots to evaluate the performance of each method under 31 degradation combinations. The results show that our method achieves the better results in both disturbance and perceptual performance.

Unified video face restoration. As shown in Table 5, we compare our method with other methods in unified video face restoration on our constructed degradation dataset. The results show that our method achieves the best results in both disturbance and perceptual performance. Visual comparisons on single frame are presented in Figure 6. Comparing with previous method, our method produces higher restoration quality while maintaining data fidelity well.

5.3. Ablation studies

Our method consists of two components, including Image-Restoration guided by prompt and Video-Refinement. In Video-Refinement component, there are two models. The model-1 is trained by low-quality images and ground-truth images in the constructed dataset. The model-2 is trained by the results of Image-Restoration and ground-truth images. The followings are the effect of these components.

Effect of Image-Restoration. As shown in Table 6, the performance of the restored face decreases on both distortion and perceptual when we remove the Image-Restoration which is guided by prompt and only retain the Video-Refinement. It indicates that the

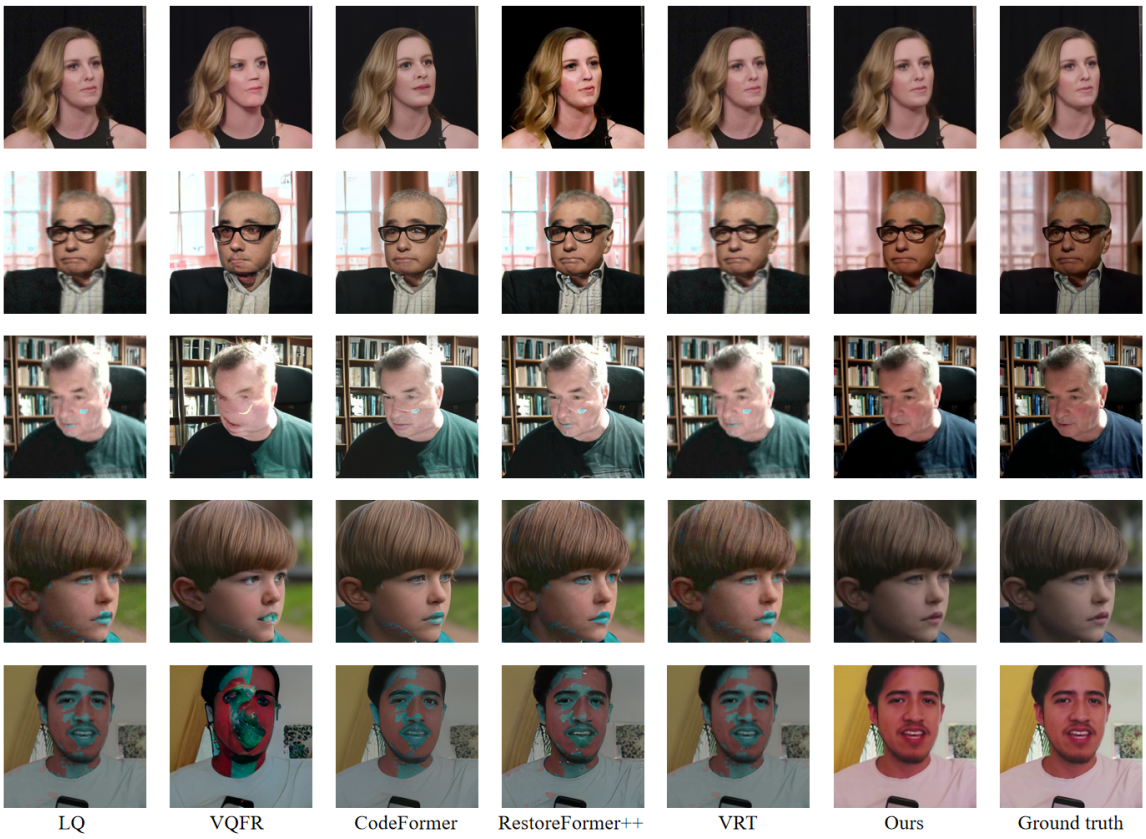

Figure 6: Visual comparisons. Our method produces higher restoration quality while maintaining data fidelity well.

Table 5: Quantitative comparison between our method with other state-of-the-art approaches on the unified face restoration task. The best results are marked in boldface.

Method	Distortion		Perceptual	
	PSNR↑	SSIM↑	LPIPS↓	FVD↓
VQFR	21.90	0.668	0.232	606.82
CodeFormer	24.22	0.731	0.209	464.56
RestoreFormer++	23.23	0.694	0.231	479.39
VRT	27.05	0.769	0.248	463.77
Ours	**32.96**	**0.916**	**0.058**	**152.06**

Image-Restoration which is guided by prompt is crucial for maintaining fidelity and improving perceptual performance.

Table 6: Ablation studies. The best results are marked in boldface.

Methods			Metrics			
Image-Restoration	Video-Refinement		Distortion		Perceptual	
	model-1	model-2	PSNR↑	SSIM↑	LPIPS↓	FVD↓
√			32.23	0.891	0.127	152.08
	√		29.69	0.891	0.101	269.66
√	√		31.88	0.910	0.059	158.11
√		√	**32.96**	**0.916**	**0.058**	**152.06**

Effect of Video-Refinement. As shown in Table 6, when we remove the Video-Refinement with model-1 which is trained by low-quality images and ground-truth images in the constructed dataset and only retain the Image-Restoration guided by prompt, the performance of the restored face images decreases in SSIM, LPIPS. Moreover, when we remove the Video-Refinement with model-2 which is trained by the results of Image-Restoration and ground-truth images, the performance of the restored face images decreases on both distortion and perceptual, which suggests that the Video-Refinement contributes to improving fidelity and achieving more perceptually appealing facial results. By comparing the effects of these two Video-Refinement models, it indicates that the model-2 trained by the results of Image-Restoration and ground-truth images is more effective.

6. Conclution

In this paper, we present a method for directing face restoration through language prompt. To broaden its applicability, we present ControlNet which eliminates the dependency on language prompt and integrates human-level knowledge from vision-language models into general networks, which enhances the performance of video face restoration and enables universal video face restoration. Furthermore, we construct a degradation dataset with multiple degradations in the same scene and corresponding face captions. It shows competitive performance in universal video face restoration.

References

Saeed Anwar, Fatih Porikli, and Cong Phuoc Huynh. Category-specific object image denoising. *IEEE Transactions on Image Processing*, page 5506–5518, Nov 2017. doi: 10.1109/tip.2017.2733739. URL http://dx.doi.org/10.1109/tip.2017.2733739.

Muhammad Asim, Fahad Shamshad, and Ali Ahmed. Blind image deconvolution using deep generative priors. *IEEE Transactions on Computational Imaging,IEEE Transactions on Computational Imaging*, Feb 2018.

Tim Brooks, Aleksander Holynski, and AlexeiA. Efros. Instructpix2pix: Learning to follow image editing instructions. Nov 2022.

T.B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, and Jared Kaplan. Language models are few-shot learners. *arXiv: Computation and Language,arXiv: Computation and Language*, May 2020.

Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. doi: 10.1109/cvpr.2018.00019. URL http://dx.doi.org/10.1109/cvpr.2018.00019.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. *Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation*, page 205–218. Jan 2023. doi: 10.1007/978-3-031-25066-8_9. URL http://dx.doi.org/10.1007/978-3-031-25066-8_9.

Ayan Chakrabarti, A.N. Rajagopalan, and Rama Chellappa. Super-resolution of face images using kernel pca-based prior. *IEEE Transactions on Multimedia*, page 888–892, Jun 2007. doi: 10.1109/tmm.2007.893346. URL http://dx.doi.org/10.1109/tmm.2007.893346.

Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K. Wong. Progressive semantic-aware style transformation for blind face restoration. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2021a. doi: 10.1109/cvpr46437.2021.01172. URL http://dx.doi.org/10.1109/cvpr46437.2021.01172.

Jia Chen, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, QuocV. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *Cornell University - arXiv,Cornell University - arXiv*, Feb 2021b.

Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. doi: 10.1109/cvpr.2018.00264. URL http://dx.doi.org/10.1109/cvpr.2018.00264.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, Jan 2019. doi: 10.18653/v1/n19-1423. URL http://dx.doi.org/10.18653/v1/n19-1423.

Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun 2019. doi: 10.1109/cvprw.2019.00232. URL http://dx.doi.org/10.1109/cvprw.2019.00232.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Oct 2020.

Nicholas D. Duran, Alexandra Paxton, and Riccardo Fusaroli. Align: Analyzing linguistic interactions with generalizable techniques—a python library. *Psychological Methods*,

page 419–438, Feb 2019. doi: 10.1037/met0000206. URL http://dx.doi.org/10.1037/met0000206.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2021. doi: 10.1109/cvpr46437.2021.01268. URL http://dx.doi.org/10.1109/cvpr46437.2021.01268.

Leonardo Galteri, Marco Bertini, Lorenzo Seidenari, Tiberio Uricchio, and Alberto Del Bimbo. Increasing video perceptual quality with gans and semantic coding. In *Proceedings of the 28th ACM International Conference on Multimedia*, Oct 2020. doi: 10.1145/3394171.3413508. URL http://dx.doi.org/10.1145/3394171.3413508.

Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder.

B.K. Gunturk, A.U. Batur, Y. Altunbasak, M.H. Hayes, and R.M. Mersereau. Eigenface-domain super-resolution for face recognition. *IEEE Transactions on Image Processing*, 12(5):597–606, May 2003. doi: 10.1109/tip.2003.811513. URL http://dx.doi.org/10.1109/tip.2003.811513.

Jingwen He, Wu Shi, Kai Chen, Lean Fu, and Chao Dong. Gcfsr: a generative and controllable face super resolution method without facial and gan priors.

Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. doi: 10.1109/iccv.2017.187. URL http://dx.doi.org/10.1109/iccv.2017.187.

Amin Jourabloo, Mao Ye, Xiaoming Liu, and Ren Liu. Pose-invariant face alignment with a single cnn. *Cornell University - arXiv,Cornell University - arXiv*, Jul 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019. doi: 10.1109/cvpr.2019.00453. URL http://dx.doi.org/10.1109/cvpr.2019.00453.

Abhinav Kumar, Tim K. Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020. doi: 10.1109/cvpr42600.2020.00826. URL http://dx.doi.org/10.1109/cvpr42600.2020.00826.

Hui Li, Zidong Guo, Seon-Min Rhee, Seungju Han, and Jae-Joon Han. Towards accurate facial landmark detection via cascaded transformers. a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. b.

Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Lin Li, and Ruigang Yang. Learning warped guidance for blind face restoration. *Cornell University - arXiv,Cornell University - arXiv*, Apr 2018.

Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and LucVan Gool. Vrt: A video restoration transformer.

Jing Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Neural Information Processing Systems,Neural Information Processing Systems*, Aug 2019.

Ziwei Luo, FredrikK Gustafsson, Zheng Zhao, Jens Sjölund, and ThomasB Schön. Controlling vision-language models for universal image restoration.

Ziwei Luo, FredrikK. Gustafsson, Zheng Zhao, Jens Sjölund, and ThomasB. Schön. Image restoration with mean-reverting stochastic differential equations. Jan 2023.

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020. doi: 10.1109/cvpr42600.2020.00251. URL http://dx.doi.org/10.1109/cvpr42600.2020.00251.

Alec Radford, JongWook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Askell Amanda, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *Cornell University - arXiv,Cornell University - arXiv*, Feb 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2022. doi: 10.1109/cvpr52688.2022.01042. URL http://dx.doi.org/10.1109/cvpr52688.2022.01042.

Joseph Roth, Tong Yiying, and Xiaoming Liu. Adaptive 3d face reconstruction from unconstrained photo collections. *IEEE Conference Proceedings,IEEE Conference Proceedings*, Jan 2016.

Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Exploiting semantics for face image deblurring. *Cornell University - arXiv,Cornell University - arXiv*, Jan 2020.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations,International Conference on Learning Representations*, Jan 2015.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Jan 2019. doi: 10.18653/v1/d19-1514. URL http://dx.doi.org/10.18653/v1/d19-1514.

Xiaoou Tang and Xiaogang Wang. Face sketch synthesis and recognition. In *Proceedings Ninth IEEE International Conference on Computer Vision*, Jan 2003. doi: 10.1109/iccv.2003.1238414. URL http://dx.doi.org/10.1109/iccv.2003.1238414.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. *Contrastive Multiview Coding*, page 776–794. Jan 2020. doi: 10.1007/978-3-030-58621-8_45. URL http://dx.doi.org/10.1007/978-3-030-58621-8_45.

Oncel Tuzel, Yuichi Taguchi, and JohnR. Hershey. Global-local face upsampling network. *Cornell University - arXiv,Cornell University - arXiv*, Mar 2016.

Georgios Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015. doi: 10.1109/cvpr.2015.7298989. URL http://dx.doi.org/10.1109/cvpr.2015.7298989.

Thomas Unterthiner, Sjoerdvan Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges. *Cornell University - arXiv,Cornell University - arXiv*, Dec 2018.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2021. doi: 10.1109/cvpr46437.2021.00905. URL http://dx.doi.org/10.1109/cvpr46437.2021.00905.

Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, page 600–612, Apr 2004. doi: 10.1109/tip.2003.819861. URL http://dx.doi.org/10.1109/tip.2003.819861.

Zhouxia Wang, Jiawei Zhang, Tianshui Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs.

Zhouxia Wang, Jiawei Zhang, Tianshui Chen, Wenping Wang, and Ping Luo. Restoreformer++: Towards real-world blind face restoration from undegraded key-value pairs. Aug 2023.

Lingbo Yang, Shanshe Wang, Siwei Ma, Wen Gao, Chang Liu, Pan Wang, and Peiran Ren. Hifacegan: Face renovation via collaborative suppression and replenishment. In *Proceedings of the 28th ACM International Conference on Multimedia*, Oct 2020. doi: 10.1145/3394171.3413965. URL http://dx.doi.org/10.1145/3394171.3413965.

Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2021. doi: 10.1109/cvpr46437.2021.00073. URL http://dx.doi.org/10.1109/cvpr46437.2021.00073.

Jianhui Yu, Hao Zhu, Liming Jiang, ChenChange Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. Mar 2023.

Xin Yu and Fatih Porikli. *Ultra-Resolving Face Images by Discriminative Generative Networks*, page 318–333. Jan 2016. doi: 10.1007/978-3-319-46454-1_20. URL http://dx.doi.org/10.1007/978-3-319-46454-1_20.

Kaipeng Zhang, Zhanpeng Zhang, Chia-Wen Cheng, Winston H. Hsu, Yu Qiao, Wei Liu, and Tong Zhang. *Super-Identity Convolutional Neural Network for Face Hallucination*, page 196–211. Jan 2018a. doi: 10.1007/978-3-030-01252-6_12. URL http://dx.doi.org/10.1007/978-3-030-01252-6_12.

Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018b. doi: 10.1109/cvpr.2018.00068. URL http://dx.doi.org/10.1109/cvpr.2018.00068.

Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Learning face hallucination in the wild. *Proceedings of the AAAI Conference on Artificial Intelligence*, Jun 2022a. doi: 10.1609/aaai.v29i1.9795. URL http://dx.doi.org/10.1609/aaai.v29i1.9795.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, page 2337–2348, Sep 2022b. doi: 10.1007/s11263-022-01653-1. URL http://dx.doi.org/10.1007/s11263-022-01653-1.

Shangchen Zhou, KelvinC.K. Chan, Chongyi Li, and ChenChange Loy. Towards robust blind face restoration with codebook lookup transformer. Jun 2022c.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and ChenChange Loy. Celebv-hq: A large-scale video facial attributes dataset. Jul 2022.