Deep Face Video Inpainting via UV Mapping

Wenqi Yang, Zhenfang Chen, Chaofeng Chen, Guanying Chen, and Kwan-Yee K. Wong, *Senior Member, IEEE*

***Abstract*—This paper addresses the problem of face video inpainting. Existing video inpainting methods target primarily at natural scenes with repetitive patterns. They do not make use of any prior knowledge of the face to help retrieve correspondences for the corrupted face. They therefore only achieve sub-optimal results, particularly for faces under large pose and expression variations where face components appear very differently across frames. In this paper, we propose a two-stage deep learning method for face video inpainting. We employ 3DMM as our 3D face prior to transform a face between the image space and the UV (texture) space. In Stage I, we perform face inpainting in the UV space. This helps to largely remove the influence of face poses and expressions and makes the learning task much easier with well aligned face features. We introduce a frame-wise attention module to fully exploit correspondences in neighboring frames to assist the inpainting task. In Stage II, we transform the inpainted face regions back to the image space and perform face video refinement that inpaints any background regions not covered in Stage I and also refines the inpainted face regions. Extensive experiments have been carried out which show our method can significantly outperform methods based merely on 2D information, especially for faces under large pose and expression variations.**

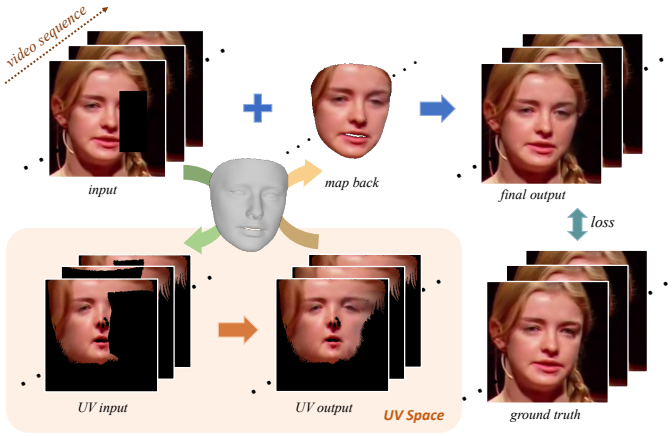


Fig. 1. Given a face video, it is preferable to learn the face texture restoration regardless of face pose and expression variations. In our proposed method, we first utilize 3D face prior (3DMM) to transform the input faces into the UV space. We then perform face inpainting in the UV space where face textures are well aligned and easy for correspondence retrieval. The restored face will then be mapped back to the image space, followed by final refinement as well as inpainting for the non-face regions to produce the final outputs.

I. Introduction

FACE video inpainting targets at restoring corrupted or occluded regions of faces in videos. It is an important research topic in computer vision and has many practical applications such as video overlay removal [1] and partially occluded face recognition in surveillance videos [2]. Note that faces in videos often exhibit diverse poses and expressions. This makes face video inpainting a challenging task.

Correspondences between frames serve as crucial clues in video inpainting for retrieving missing information from neighboring frames and ensuring temporal consistency. Existing video inpainting methods mainly focus on restoring the backgrounds of natural scenes which are mostly stationary and consist of repetitive patterns. They typically fill missing regions by copying and propagating similar patterns or textures from other regions [3]–[6]. However, directly referring to other frames often results in improper contents when elements in a video move around and change their appearances. Hence, these methods are only capable of tackling narrow masks and static background. Recently, a number of learning-based methods have been proposed [1], [7]–[13]. These methods successfully learn domain knowledge from an enormous number of training samples and can generate proper content for large missing regions. Most of them are based on spatial-temporal attention or assisted by optical flow to learn the correspondences across frames, and are suitable for natural scenes with simple motions. For face videos, however, the appearance of a face can vary a lot under different poses and expressions. These methods have difficulties in finding proper reference in neighboring frames to restore reasonable contents for faces. They often fail to generate visually plausible face structures when no reference can be found in neighboring frames due to their lack of face prior knowledge. Hence, they cannot guarantee recovering proper faces in the videos.

Owing to the prior knowledge of the 3D face structure, human can interpret, recognize, or even "reconstruct" a corrupted face image with relative ease. For instance, human can recognize faces in low quality videos under diverse viewpoints and partial occlusions, as well as under different face poses and expressions. Inspired by this, we propose to exploit 3D face prior for face video inpainting. In this paper, we employ an expressive 3D face model as our 3D face prior. By fitting this 3D face model to the video frames, we can transform the face from the image space to the UV (texture) space and vice

Wenqi Yang and Kwan-Yee K. Wong are with the Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong. (E-mail: [wqyang, kykwong]@cs.hku.hk)

Zhenfang Chen is with MIT-IBM Watson AI Lab, Cambridge, MA, USA, 02142 (E-mail: chenzhenfang2013@gmail.com).

Chaofeng Chen is with Nanyang Technological University, Singapore (E-mail: chaofenghust@gmail.com).

Guanying Chen is with FNii and SSE, The Chinese University of Hong Kong, Shenzhen, China (E-mail: chenguanying@cuhk.edu.cn).

versa. Note that faces in the UV space represent unwarped face textures which are well aligned. This helps to remove the influence of poses and expressions, and makes the learning of face structure much easier. Besides, the well alignment and symmetry of the face features in the UV space also make it trivial to locate correspondences in neighboring frames which provide rich information for face video inpainting. Based on these observations, we propose to carry out face video inpainting in the UV space (see Fig. 1). We introduce a Multi-reference UV-map Completion Network (MUC-Net) with a novel Frame-wise Attention (FA) module to perform reference-based face completion in the UV space.

Our proposed method is a two-stage approach. As a pre-processing step, we fit the 3D Morphable Model (3DMM) [14] to every frame of the face video. We use the estimated model parameters to transform the face between the image space and the UV space in the two core stages. In Stage I, namely UV-map completion stage, we first transform the face to the UV space and carry out UV-map completion using our proposed MUC-Net. Our FA module is designed specifically to take full advantage of the well-aligned face features in the UV space to find proper correspondences in neighboring frames in an efficient and effective manner. In Stage II, namely face video refinement stage, we transform the inpainted UV-map back to the image space and perform face video refinement using our proposed Face Video Refinement Network (FVR-Net). FVR-Net inpaints any background regions not covered in Stage I and at the same time refines the inpainted face regions.

In contrast to other methods, our method ensures the plausibility of face structure through the use of 3D face prior. Our method is more robust for faces under large pose and expression variations, and can better exploit correspondences in neighboring frames. Our key contributions include:

- To the best of our knowledge, we are the first to perform face video inpainting via the UV space. Thanks to the well alignment and symmetry of the face features in the UV space, our MUC-Net can robustly restore the missing face regions with plausible face structures and textures.
- We propose a novel Frame-wise Attention (FA) module that can take full advantage of the well aligned face features in the UV space to find proper correspondences efficiently in neighboring frames to assist face inpainting.
- Our method achieves state-of-the-art performance in face video inpainting, especially for the challenging cases with large face pose and expression variations. Comprehensive experiments demonstrate the effectiveness and robustness of our method.

II. Related Work

A. Face Image Inpainting

Traditional image inpainting methods fill the missing regions by progressively propagating pixels from the neighboring regions [15]–[17] or by iteratively searching for matching patches [18]–[22]. These conventional methods are capable of handling cases with stationary textures and relatively small masked regions, but fail when textures and structures are non-repetitive. Some studies [23], [24] exploit specific face domain knowledge for face image inpainting. Constrained by the representation ability of their models, however, they can only restore specific face regions in frontal faces.

Recently, learning-based methods [25]–[35] have been proposed to perform inpainting by learning from large image datasets. These methods are more robust and expressive than the traditional non-learning-based methods. Some of them focus on improving the structure and texture coherence by introducing additional guidance such as edges [36], segmentation [37], structure flow [38], and foreground contour [39]. Others [40]–[42] explore feature matching between masked regions and known regions in the semantic space by proposing new modules such as contextual attention [43]. There are also inpainting works that focus on progressively filling in the masked regions from the their boundaries and treating masked regions (usually input as 0 value) and non-masked regions separately, such as partial convolution [44] and gated convolution [45]. Partial convolution [44] updates a binary mask regularly, while gated convolution [45] adopts a learnable soft valued mask.

In response to the need of face related applications, a number of methods [23], [24], [46] have been proposed specifically for the face inpainting task. To stabilize the restored face structures, some of them introduce additional guidance such as landmarks [47], [48] and face parsing [49], [50] into the pipeline to serve as intermediate output or loss term. Others propose to make use of additional inputs such as reference images [51] from the same person to preserve identities, or colorized sketches [52], [53] to perform interactive face editing (modify the shape / color of the given face). Li *et al*. [54] propose to utilize face symmetry by performing illumination-aware feature warping from flipped images. However, human face may not look strictly symmetric under large pose variations. Another batches of works [55]–[59] focus on blind face image restoration without masks indicating corrupted regions. Based on observations, most of the face image inpainting solutions perform unsatisfactorily under large pose or expression variations. This is due to their lack of 3D face priors to help understand and restore face structures from 2D images. Furthermore, these image-based methods can only achieve sub-optimal results in face video inpainting as they do not exploit information provided by correspondences in neighboring frames.

B. Video Inpainting

Different from image inpainting, video inpainting takes a sequence of frames as input and restores the missing regions based on both spatial and temporal information. Compared with image-based methods, video inpainting methods explore correspondences between frames as crucial clues to retrieve missing information from neighboring frames to ensure temporal consistency. Traditional video inpainting methods [3], [4], [60], [61] typically perform patch-based or optical-flow-based optimizations which require heavy computations. They are capable of generating plausible contents for general videos consisting of stationary background with repetitive patterns and consistent textures. However, they may fail miserably

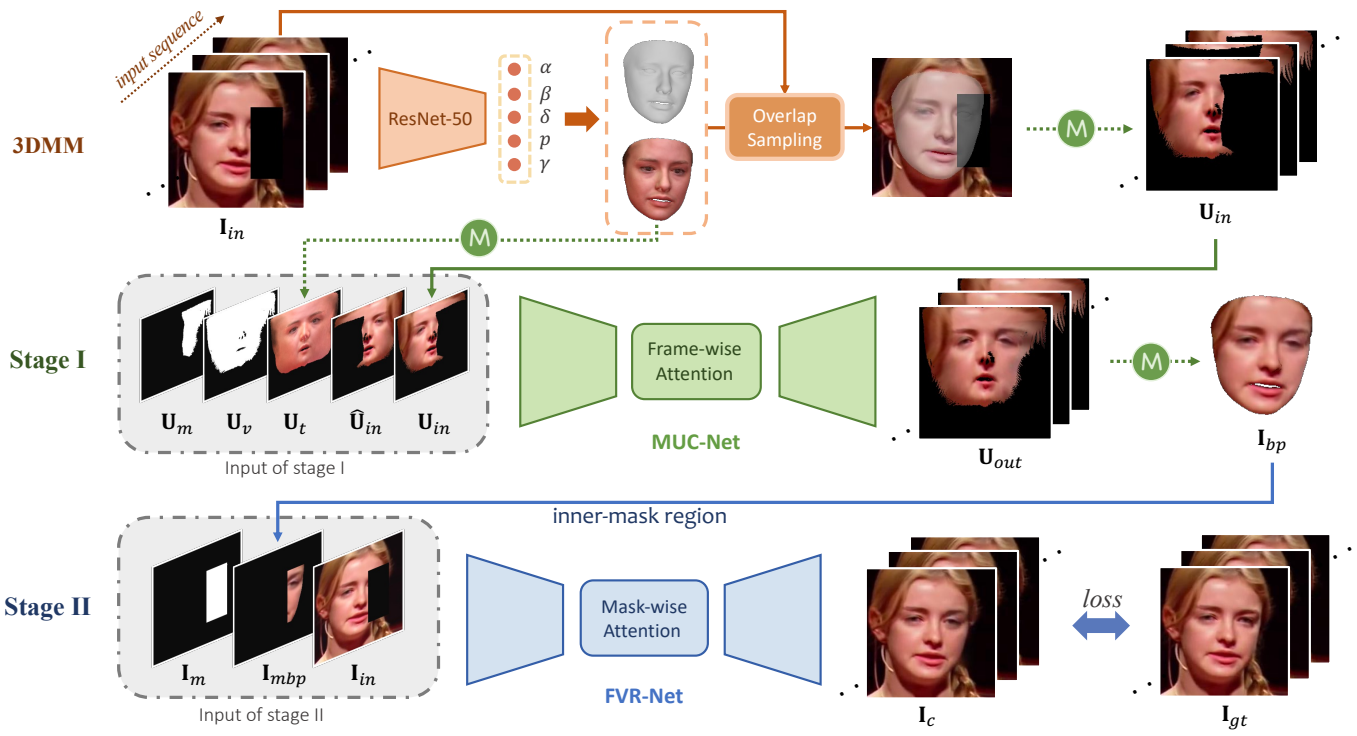


Fig. 2. The overall pipeline of our method, where "Ⓜ" represents UV mapping. Our framework consists of two stages, namely (I) UV map completion and (II) Face video refinement. Given a face video sequence, we first utilize 3DMM as our face prior to predict the face structures by model regression. In Stage I, we transform the face texture from image space to UV space via the fitted 3DMM, and perform UV map completion to get preliminary results. In Stage II, we take the mapped-back face together with the corrupted face to generate the final output, where both face regions and background (non-face regions) will be refined or inpainted.

when structures and textures are complicated and their appearances vary largely across frames. Boosted by deep learning techniques, learning-based methods have been proposed to explore solutions for better utilizing spatial and temporal information by introducing flow-warping [8]–[10], [13], [62], [63], cross-frame attention [11], [64], and 3D convolution [1], [7], [12]. Optical flow is often adopted as an intermediate guidance [13], [62] or used to warp frames into alignment [63]. This facilitates the calculation of warping loss [9].

The above methods mainly retrieve correspondences by searching for similar patches or making the patterns aligned based on flow-warping. However, human face can appear very differently under large pose and expression variations. This makes it more difficult to find proper reference from neighboring frames due to the large appearance variations. There is also a video inpainting work [65] focusing on face re-identification. They target at the restoration of the de-identified face videos by taking the original face landmarks as input. The mask is designed to cover all the key face components for all the input frames while the background is preserved. Under this setting, no reference is available from other frames to recover the masked regions. They instead focus on predicting consistent identity for all the frames from the given landmarks. In this paper, we aim at efficiently retrieving proper correspondences from neighboring frames for face video inpainting. We exploit an effective way to transform face textures into a well-aligned space which greatly facilitates both correspondence learning and feature restoration.

C. 3D Face Prior

Human domain knowledge has become a powerful tool in numerous tasks owing to the learnable human prior (e.g., body structure [66] and face prior [67]). In this paper, we focus on face prior assisted face video inpainting. Commonly used face priors include face parsing, face landmarks, and face model [68], [69]. In particular, 3D face morphable model (3DMM) [14], [70], [71], [71]–[73] has achieved stable and excellent performance in face reconstruction, and has been wildly adopted in face related works such as face recognition [74], face frontalization [75], face editing [76], makeup transfer [77], face reenactment [78], face super-resolution [79], face deblurring [80], and animation [81]. The impressive results of these works well demonstrate the advantages of embedding 3D face prior into face-related tasks.

Among works that utilize 3D face prior, UV-GAN [74] is closely related to our work. UV-GAN also utilizes face model and UV map to recover face regions. However, their motivation and contributions are different from ours. UV-GAN is proposed to reconstruct face models and synthesize novel views to enlarge the diversity of poses for pose-invariant face recognition. They exploit UV textures and leverage symmetry of the face to recover self-occluded regions in the fitted model. They only deal with single images. In this paper, we target at robust face video inpainting by making use of 3D face prior to facilitate both face structure learning and correspondence finding from the well-aligned feature maps in the UV space.

III. METHOD

A. Overview

As briefly introduced in Sec. I, our method is a two-stage approach. Fig. 2 shows an overview of our proposed pipeline. Given a face video, we first fit 3DMM to every frame to obtain per-frame shape, texture, and pose parameters. The shape and pose parameters are used for transforming the face between the image space and UV space, while the texture parameters are used to generate synthesized texture to provide auxiliary information for the inpainting task. In Stage I, we first transform the face from the image space to the well-aligned UV space and use our proposed MUC-Net to perform UV-map completion. FA module is proposed for MUC-Net to facilitate the correspondence retrieval across UV texture frames. In Stage II, we transform the completed UV-map in Stage I back to the image space, and use our proposed FVR-Net to inpaint any background (non-face) regions not covered in Stage I as well as refine the inpainted face regions.

In Sec. III-B, we will first give a brief review of 3DMM which serves as our 3D face prior and facilitates the transformation of faces between the image space and the UV space. We will then describe the details of our MUC-Net and FA module for UV-map completion in III-C. Details of face video refinement using our FVR-Net will be covered in Sec. III-D .

B. 3D Face Prior

1) Face Reconstruction: In this work, we employ 3DMM [14] as our 3D face prior. We adopt the method proposed by Deng *et al*. [70] to fit 3DMM to the video frames using a modified ResNet-50 network [82]. We retrain the network with masked face images as input, and the output is a combined vector $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{p}) \in \mathbb{R}^{257}$, where $\boldsymbol{\alpha} \in \mathbb{R}^{80}$, $\boldsymbol{\beta} \in \mathbb{R}^{64}$, $\boldsymbol{\delta} \in \mathbb{R}^{80}$, $\boldsymbol{\gamma} \in \mathbb{R}^{27}$, and $\boldsymbol{p} \in \mathbb{R}^{6}$ represent face identity, expression, texture, illumination, and pose respectively. Concretely, the pose vector $\boldsymbol{p}$ is composed of a rotation vector[1] $\boldsymbol{r} \in \mathbb{R}^3$ and a translation vector $\mathbf{t} \in \mathbb{R}^3$. With the predicted parameters, the shape $\mathbf{S}$ and texture $\mathbf{T}$ of the 3D face can be modeled as:

$$\begin{aligned} \mathbf{S} &= \bar{\mathbf{S}} + \mathbf{B}_{id}\boldsymbol{\alpha} + \mathbf{B}_{exp}\boldsymbol{\beta}, \\ \mathbf{T} &= \bar{\mathbf{T}} + \mathbf{B}_{tex}\boldsymbol{\delta}, \end{aligned} \tag{1}$$

where $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$ denote the mean shape and texture, $\mathbf{B}_{id}$, $\mathbf{B}_{exp}$, and $\mathbf{B}_{tex}$ are the PCA bases for face identity, expression, and texture respectively. Similar to Deng *et al*. [70], we adopt $\bar{\mathbf{S}}$, $\bar{\mathbf{T}}$, $\mathbf{B}_{id}$, and $\mathbf{B}_{tex}$ from BFM [83], and $\mathbf{B}_{exp}$ built from FaceWarehouse [84].

2) Texture Sampling and UV Mapping: Given the predicted shape $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and pose $(\boldsymbol{r}, \boldsymbol{t})$ parameters, we can transform a face from the image space to the UV space through texture sampling and UV mapping. We first project the 3D face model onto the image using the pose parameters and perform bilinear sampling to compute per-vertex texture for the 3D face model. For self-occluded and back-facing vertices, as well as vertices projected onto the masked regions, we simply assign zero to their texture values. Finally, we carry out UV mapping to transform the 3D face model texture to the UV space.

[1]Euler angles *yaw*, *pitch*, and *roll* for constructing the rotation matrix $\mathbf{R} \in \mathrm{SO}(3)$

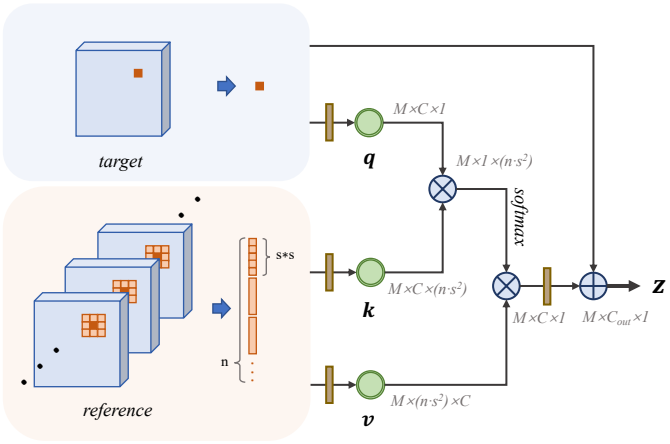


Fig. 3. Detailed structure of our Frame-wise Attention (FA) module. The symbol "⊗", "⊕" in the figure represent multiplication and element-wise sum, respectively. M is the number of points in the masked regions (embedding space); C is the channel dimension; $n \cdot s^2$ gives the number of selected memory points (key and value) from the reference frames.

For the rest of this paper, we denote a corrupted input frame and its ground truth as $\mathbf{I}_{in}$ and $\mathbf{I}_{gt}$ respectively, and their UV-maps as $\mathbf{U}_{in}$ and $\mathbf{U}_{gt}$ respectively. We represent the missing regions in the image space using a 2D binary mask $\mathbf{I}_m$, and denote its UV-map as $\mathbf{U}_m$. Similarly, we represent the valid projection of the 3D face model in the image space using a 2D binary mask $\mathbf{I}_v$, and denote its UV-map as $\mathbf{U}_v$ (see Fig. 2). We also map the synthesized texture $\mathbf{T}$ to the UV space and denote it as $\mathbf{U}_t$.

C. Stage I: UV-map Completion

We first transform the face from the image space to the UV space and carry out UV-map completion. As mentioned previously, the UV maps of a face represent unwarped face textures which are well aligned and largely invariant to face poses and expressions. This greatly facilitates the learning of face structures and the finding of correspondences in neighboring frames.

1) Multi-reference UV-map Completion Network (MUC-Net): We adopt an encoder-decoder network equipped with gated convolutions [45] as the backbone of our MUC-Net (network details can be found in the supplementary material). We concatenate each frame $\mathbf{U}^i_{in}$ with its horizontally flipped UV-map $\hat{\mathbf{U}}^i_{in}$, synthesized texture map $\mathbf{U}^i_t$, valid face projection $\mathbf{U}^i_v$, and missing regions $\mathbf{U}^i_m$, and feed them to the encoder to generate the feature map $\mathbf{F}^i$:

$$\mathbf{F}^i = En(\mathbf{U}^i_{in}, \hat{\mathbf{U}}^i_{in}, \mathbf{U}^i_t, \mathbf{U}^i_v, \mathbf{U}^i_m). \tag{2}$$

The flipped UV map $\hat{\mathbf{U}}_{in}$ exploits symmetry to provide auxiliary information when only parts of the symmetrical face features are being masked, whereas the synthesized texture map $\mathbf{U}_t$ helps to provide auxiliary information when symmetrical face features are being completely masked.

To exploit information provided by correspondences in neighboring frames, we propose a Frame-wise Attention (FA)

module to fuse features from neighboring frames. Specifically, for each *target frame*, we select n other frames as its *reference frames* and fuse their features using the FA module:

$$\mathbf{Z}^i = Attn(\mathbf{F}^i, \{\mathbf{F}^{i+j} \mid j \in \Omega\}), \tag{3}$$

where Ω is the set of offset indices for the reference frames. In our experiments, we take $\Omega = \{-2, -1, +1, +2\}$. Finally, the fused feature map $\mathbf{Z}^i$ is fed to the decoder to generate the completed UV map $\mathbf{U}^i_{out}$:

$$\mathbf{U}^i_{out} = De(\mathbf{Z}^i). \tag{4}$$

2) Frame-wise Attention: Inspired by the recently proposed attention mechanism [85], we design a frame-wise attention block to explore correspondences between a target frame and its reference frames. Thanks to the well alignment of the face features in the UV space, we can limit our search for correspondences in a small local window. Concretely, we pick *query* points from the masked regions in the feature map of the target frame. For each *query*, we define a $s \times s$ small window (we set $s = 3$ in our experiments) centered at the *query* for selecting its *reference* points from the feature maps of the reference frames (see Fig. 3). This small window design is employed to account for any slight misalignment of the UV maps. Given the query $\mathbf{q} \in \mathbb{R}^C$ evaluated at the query point, and the keys $\mathbf{K} \in \mathbb{R}^{C\times(s^2\times n)}$ and values $\mathbf{V} \in \mathbb{R}^{C\times(s^2\times n)}$ evaluated at the reference points, frame-wise attention is accomplished by

$$\begin{aligned} \boldsymbol{\alpha} &= \frac{\exp\left(\mathbf{K}^{\mathrm{T}}\mathbf{q}\right)}{\sum_{m=1}^{N}\exp\left(\mathbf{K}_m^{\mathrm{T}}\mathbf{q}\right)}, \\ \mathbf{z} &= \mathbf{f} + W_z(\mathbf{V}\boldsymbol{\alpha}), \end{aligned} \tag{5}$$

where $N = s^2 \times n$ gives the total number of reference points; $\boldsymbol{\alpha}$, $\mathbf{f}$, $\mathbf{z}$, and W_z denote the attention vector, input feature vector, output feature vector, and output embedding layers respectively. Compared with previous works such as STTN [86] which uses spatial-temporal non-local attention to find correspondences across different frames, our design dramatically cuts down unnecessary computations and greatly improves the time complexity.

3) Loss Functions: We use $\mathcal{L}_1$ loss and SSIM loss [87] for both UV maps and back-projected faces to train MUC-Net. $\mathcal{L}_1$ loss aims at minimizing the distance between the ground-truth and predicted UV maps, whereas SSIM loss is adopted for maximizing structure similarity. The loss for the UV map is computed as

$$\begin{aligned} \mathcal{L}_U &= \mathcal{L}^U_1 + \mathcal{L}^U_{SSIM}, \\ \mathcal{L}^U_1 &= \left\|\mathbf{U}_v \circ (\mathbf{U}_{out} - \mathbf{U}_{gt})\right\|_1, \\ \mathcal{L}^U_{SSIM} &= -\frac{\left(2\mu_{\mathbf{U}_{out}}\mu_{\mathbf{U}_{gt}} + c_1\right)\left(2\sigma_{\mathbf{U}_{out}\mathbf{U}_{gt}} + c_2\right)}{\left(\mu^2_{\mathbf{U}_{out}} + \mu^2_{\mathbf{U}_{gt}} + c_1\right)\left(\sigma^2_{\mathbf{U}_{out}} + \sigma^2_{\mathbf{U}_{gt}} + c_2\right)}, \end{aligned} \tag{6}$$

where $\circ$ denotes element-wise multiplication, μ and σ denote the mean and variance, c_1 and c_2 are stabilization constants set as 0.01^2 and 0.03^2 according to [87]. Similarly, we define $\mathcal{L}^I_1$, $\mathcal{L}^I_{SSIM}$, and $\mathcal{L}_I$ for the back-projected face $\mathbf{I}_{bp}$ in the image space. The overall loss for Stage I is given by

$$\mathcal{L}_{(\mathrm{I})} = \lambda_\alpha \cdot \mathcal{L}_U + \lambda_\beta \cdot \mathcal{L}_I, \tag{7}$$

where the weights λ_α and λ_β are empirically set as 1.0 and 2.0 respectively.

D. Stage II: Face Video Refinement

We transform the output from MUC-Net back to the image space by rendering the 3D face model with the predicted UV map, and denote this back-projected face as $\mathbf{I}_{bp}$. We then perform face video refinement to inpaint any background (non-face) regions not covered in Stage I as well as to refine and fuse the inpainted face regions with the input frame.

1) Face Video Refinement Network (FVR-Net): Similar to MUC-Net, we adopt an encoder-decoder network as the backbone for our FVR-Net (networks details can be found in the supplementary material). We concatenate each frame $\mathbf{I}^i_{in}$ with its masked back-projected face $\mathbf{I}^i_{mbp} = \mathbf{I}^i_m \circ \mathbf{I}^i_{bp}$ and missing regions $\mathbf{I}^i_m$, and feed them to the encoder to generate the feature map $\tilde{\mathbf{F}}^i$. A Mask-wise Attention (MA) module is proposed to fuse features from non-masked regions in neighboring frames. MA block is similar to the FA block, but with the reference points taken from the non-masked regions of both the target and reference frames. The fused feature map is fed to the decoder to generate the predicted image $\mathbf{I}_{out}$. The final output $\mathbf{I}^i_c$ is then obtained by

$$\mathbf{I}^i_c = \mathbf{I}^i_m \circ \mathbf{I}^i_{out} + (1 - \mathbf{I}^i_m) \circ \mathbf{I}^i_{in}. \tag{8}$$

2) Loss Functions: Similar to Stage I, we adopt $\mathcal{L}^I_{SSIM}$ and a slightly modified version of $\mathcal{L}^I_1$ to train FVR-Net. In addition, we also use perceptual loss $\mathcal{L}^I_{per}$ to minimize the distance in the semantic feature space. The overall loss for Stage II is given by

$$\mathcal{L}_{(\mathrm{II})} = \mathcal{L}^I_{1+} + \mathcal{L}^I_{SSIM} + 0.1 \cdot \mathcal{L}^I_{per}, \tag{9}$$

where

$$\begin{aligned} \mathcal{L}^I_{1+} &= \left\|\mathbf{I}_{out} - \mathbf{I}_{gt}\right\|_1 + 2 \cdot \left\|\mathbf{I}_m \circ (\mathbf{I}_{out} - \mathbf{I}_{gt})\right\|_1, \\ \mathcal{L}^I_{SSIM} &= -\frac{\left(2\mu_{\mathbf{I}_{out}}\mu_{\mathbf{I}_{gt}} + c_1\right)\left(2\sigma_{\mathbf{I}_{out}}\sigma_{\mathbf{I}_{gt}} + c_2\right)}{\left(\mu^2_{\mathbf{I}_{out}} + \mu^2_{\mathbf{I}_{gt}} + c_1\right)\left(\sigma^2_{\mathbf{I}_{out}} + \sigma^2_{\mathbf{I}_{gt}} + c_2\right)}, \\ \mathcal{L}^I_{per} &= \frac{1}{C_k H_k W_k}\left\|\phi_k\left(\mathbf{I}_{out}\right) - \phi_k\left(\mathbf{I}_{gt}\right)\right\|_2^2, \end{aligned} \tag{10}$$

where ϕ_k is the k-th layer output of a pretrained VGG-16 network [91], C_k, H_k, and W_k denote the channel number, height, and width of the k-th layer output respectively.

IV. Experiments

A. Implementation Details

1) Dataset: We use the 300VW [92] dataset for our experiments. 300VW dataset contains 114 face videos with diverse face poses and expressions. We excluded low quality videos and selected 75 videos for training and 20 for evaluation.

2) Inpainting Settings: We followed the pre-processing described in Deng *et al.* [70] to crop and resize the face regions. The image size adopted for face videos is 224×224 and the UV maps have a dimension of 256×256. To verify our contribution in handling large pose variations, we extracted every 10-th frames from the original face videos as our test sequences. Since our method is robust to pose variations, we randomly sample reference frames from the whole sequence for the training phase.

TABLE I
Comparison with the state-of-the-art video inpainting methods.

Method	Rectangular Mask								Irregular Mask							
	Shifting Mask				Static Mask				Shifting Mask				Static Mask			
	$\ell_1 \downarrow$	PSNR↑	SSIM↑	VFID↓	$\ell_1 \downarrow$	PSNR↑	SSIM↑	VFID↓	$\ell_1 \downarrow$	PSNR↑	SSIM↑	VFID↓	$\ell_1 \downarrow$	PSNR↑	SSIM↑	VFID↓
DeepfillV2 [45]	0.0624	21.22	0.9470	0.2826	0.0665	20.66	0.9569	0.2751	0.0665	20.74	0.9358	0.5649	0.0567	22.43	0.9694	0.3847
LaFin [47]	0.0658	21.19	0.9450	0.2975	0.0724	20.41	0.9529	0.3283	0.0647	21.35	0.9400	0.3701	0.0597	22.37	0.9693	0.3036
STN-GAN [65]	0.0644	21.23	0.9505	0.2547	0.0686	20.57	0.9573	0.3153	0.0574	21.90	0.9494	0.3120	0.0524	23.14	0.9745	0.2330
Co-Mod-GAN [88]	0.0784	20.16	0.9382	0.4341	0.0780	20.11	0.9519	0.4003	0.0702	20.85	0.9398	0.4298	0.0634	22.07	0.9691	0.3358
MAT [89]	0.0782	19.75	0.9345	0.4046	0.0937	18.24	0.9402	0.6562	0.0739	20.06	0.9309	0.3960	0.0636	21.46	0.9660	0.2840
VINet [9]	0.0779	21.72	0.9556	0.1772	0.1354	17.34	0.9402	0.5028	0.0887	20.30	0.9461	0.2246	0.1463	18.03	0.9516	0.3749
3DGated [1]	0.0436	24.40	0.9658	0.1874	0.0663	20.82	0.9574	0.4170	0.0437	24.40	0.9598	0.2757	0.0510	23.48	0.9729	0.2699
STTN [86]	0.0385	25.18	0.9684	0.1829	0.0571	21.89	0.9627	0.3372	0.0389	24.94	0.9655	0.2235	0.0484	23.49	0.9756	0.2213
FuseFormer [90]	0.0508	22.98	0.9558	0.2745	0.0537	22.52	0.9647	0.2685	0.0491	23.03	0.9529	0.3265	0.0424	24.71	0.9778	0.2102
Ours	**0.0372**	**25.36**	**0.9709**	**0.1587**	**0.0466**	**23.64**	**0.9710**	**0.2649**	**0.0342**	**25.92**	**0.9706**	**0.1813**	**0.0344**	**26.41**	**0.9832**	**0.1816**

Rectangular Mask (from top to bottom: case A, B, C, D)

Irregular Mask (from top to bottom: case A, B, C, D)

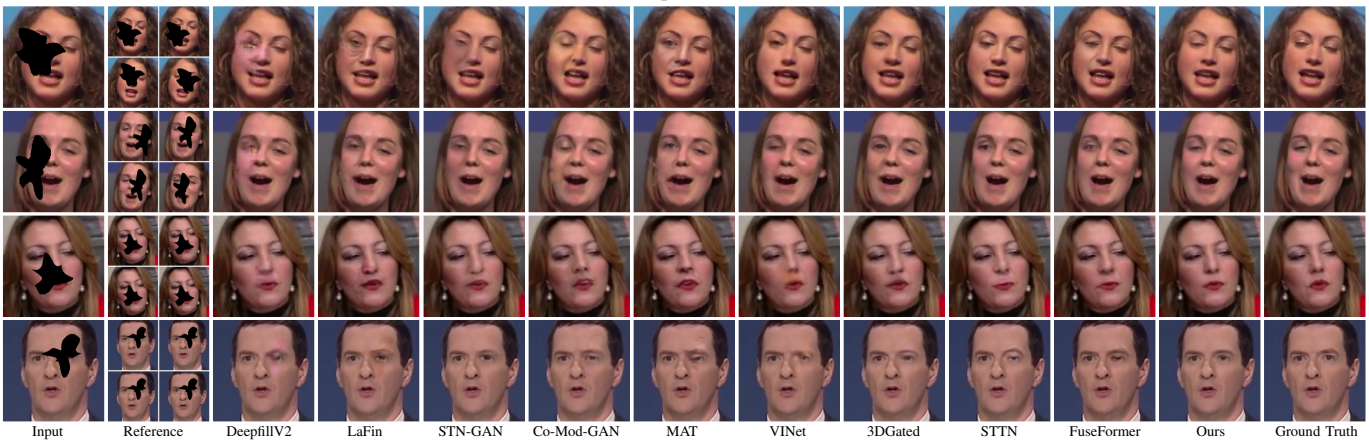


Fig. 4. Comparison with the state-of-the-arts. The top four columns are rectangular mask cases and bottom four are irregular mask cases.

TABLE II
Quantitative analysis on effectiveness of UV-map completion. Here "singles" means single frame and "multi" means multi-frame.

Method	Shifting Mask				Static Mask			
	$\ell_1 \downarrow$	PSNR ↑	SSIM ↑	VFID ↓	$\ell_1 \downarrow$	PSNR ↑	SSIM ↑	VFID ↓
single, w/o $\mathbf{I}_{mbp}$	0.0544	22.53	0.9555	0.2547	0.0586	21.94	0.9627	0.3293
single	0.0513	23.12	0.9586	0.2425	0.0541	22.78	0.9660	0.2889
multi, w/o $\mathbf{I}_{mbp}$	0.0395	24.89	0.9676	0.1848	0.0559	22.04	0.9638	0.3917
Ours	**0.0372**	**25.36**	**0.9709**	**0.1587**	**0.0466**	**23.64**	**0.9710**	**0.2649**

Fig. 5. Qualitative analysis on effectiveness of UV-map completion.

3) Mask Settings: We consider two kinds of masks for evaluation:

- **Shifting masks** are generated with slightly altered shapes and quick motions across frames, which mimic non-stationary occlusions in face videos.
- **Static masks** keep consistent shapes and locations for the whole video sequence, which also commonly happen in real scenes.

We also consider two kinds of mask shapes:

- **Rectangular masks** are a representative case which is commonly used in inpainting tasks.
- **Irregular masks** [86] mimic arbitrarily shaped occlusion objects in face videos.

The generated masks occupy between 8%-20% of the whole image. Irregular masks are only evaluated in the baseline comparison (Sec. IV-B). We test both mask shapes under the shifting and static cases.

4) Metric Settings: We consider four different metrics in our quantitative evaluations, namely (1) ℓ_1 error; (2) PSNR (Peak Signal-to-Noise Ratio); (3) SSIM [87] (Structural Similarity); and (4) VFID [1], [93] (Video-based Fréchet Inception Distance, a video perceptual measure).

B. Comparison with State-of-the-Arts

In this section, we conducted comparison between the proposed method and other inpainting methods to illustrate the strength of our framework.

1) Baselines: To the best of our knowledge, limited works have been proposed for face videos and consider the combination of face prior with video inpainting pipeline. We therefore look for video inpainting works that have been tested on face videos with codes publicly available, and select [1] for comparison. Apart from [1], we also select two representative video inpainting baselines [9], [86] and two image inpainting works [45], [47] for comparison. To evaluate the effectiveness of our method on face videos, we also reimplement a face video re-identification method [65] for comparison. All the baselines are recent deep learning methods developed for general scenes or face images. Below gives a brief summary of them:

- **DeepfillV2** [45], an encoder-decoder structured method based on 2D gated convolutions and contextual attention.
- **LaFin** [47], a landmark-guided two-stage method proposed for face image inpainting.
- **STN-GAN** [65], a GAN-based model proposed for face re-identification using 3D Residual blocks to aggregate features.
- **Co-Mod-GAN** [88], a GAN-based image inpainting method with co-modulation of both conditional and stochastic style representations.
- **MAT** [89], a transformer-based model for large hole image inpainting.
- **VINet** [9], a context aggregation method based on recurrent structures and flow-warping.
- **3DGated** [1], an encoder-decoder network based on 3D gated convolutions.
- **STTN** [86], a transformer-based method by spatial and temporal patch-matching.
- **FuseFormer** [90], a transformer-based method with fine-grained feature fusion for video inpainting.

For a fair comparison, we retrained their models on the same dataset using their publicly-available code. Since codes are not available for STN-GAN [65], we reimplemented their method according to the details in their paper. However, due to the nature of their task, they require ground truth landmarks as additional inputs, which are not available for face inpainting tasks. We therefore trained a landmark prediction network [94] to predict landmarks from the corrupted faces for them. Since the model design of Co-Mod-GAN [88] and MAT [89] requires the resolution of input images to be a power of 2, We resize the input images to 256. Therefore, quantitative results may be affected to some extend.

2) Quantitative Comparison: Table I summarizes the quantitative comparison results, where our method consistently outperformed the other methods on all four metrics under the two different mask settings. Due to the lack of face priors, all three video baselines fail to reconstruct the faces under the static mask setting. Due to the difficulty of correspondence retrieval in face videos with large pose / expression variations, they also performed poorly in the shifting mask setting. For image-based methods, even though face prior was utilized, they still failed since no temporal information was considered. Further, it is observed that the performance of landmark guided method may be affected by the limited accuracy of the landmarks predicted from corrupted faces.

3) Qualitative Comparison: We further conducted visual comparison on four classic scenes:

(A) Face expression appears differently in reference frames;
(B) Face pose changes frequently;
(C) No useful reference in other frames (*e.g.*, static masks);
(D) No useful reference in other frames, however, it can be self-referenced (*e.g.*, one eye covered).

Results are shown in Fig. 4 for both rectangular masks and irregular masks. For each kind of masks, case A, B, C, and D are presented from top to bottom.

Since we target at face videos, where correspondence retrieval is much more difficult than general scenes due to large face pose and expression variations, all the video baselines failed in these challenging cases (A & B). Specifically, in case A, other video-based methods either attended to or

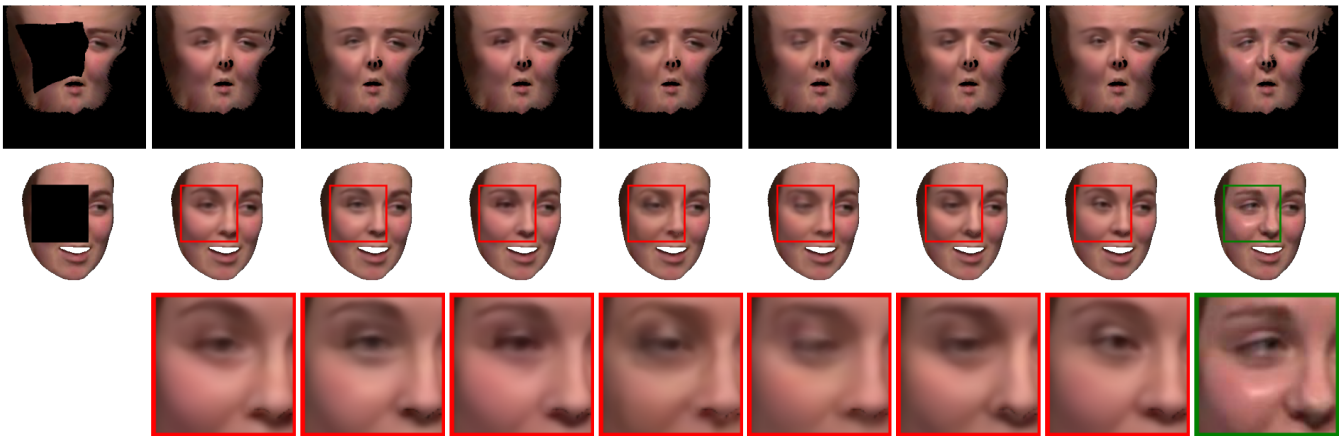


Fig. 6. Qualitative results of the ablation study on UV-map completion. Here from top to bottom: UV map, back-projected face, and the highlighted region.

TABLE III
Analysis of UV-map completion stage. Here "single" means trained with single frame as input; "concat" means using concatenated frames as input; "fuse" means fusing (concatenating) the feature maps of all $n+1$ frames before the decoder.

Method	Shifting Mask				Static Mask			
	$\ell_1 \downarrow$	PSNR ↑	SSIM ↑	VFID ↓	$\ell_1 \downarrow$	PSNR ↑	SSIM ↑	VFID ↓
w/o $\hat{\mathbf{U}}_{in}$	0.0399	24.68	0.9723	0.2042	0.0495	23.27	0.9726	0.2976
w/o $\mathbf{U}_t$	0.0397	24.70	0.9728	**0.1981**	0.0505	23.05	0.9725	0.2887
w/o $\mathbf{U}_v$	0.0400	24.72	0.9726	0.2058	0.0496	23.33	0.9729	0.2901
single	0.0520	22.93	0.9639	0.2724	0.0542	22.73	0.9699	0.3095
concat	0.0459	23.84	0.9679	0.2416	0.0538	22.79	0.9701	0.3121
fuse	0.0420	24.36	0.9708	0.2077	0.0509	23.11	0.9716	0.2991
Ours	**0.0394**	**24.80**	**0.9731**	0.2007	**0.0483**	**23.47**	**0.9735**	**0.2824**

directly copied the opened eyes from the reference frames and produced incorrect results. In case B, when face pose varied largely between frames, even though reference could be retrieved from other frames, they failed to comprehend the 3D face structure and directly incorporated the nose under a different pose to the target frame, resulting in a distorted face.

For case C and case D, due to the lack of face prior, they all failed to predict proper face structures when no useful reference could be obtained (though it could be self-referenced in case D). The flow-based context aggregation method VINet failed completely in the static mask setting. As expected, our method performed the best on these challenging cases and achieved the most visually pleasant results compared to the other baselines. Through the use of 3D face prior, our method can take full advantage of the well alignment and symmetry properties of the UV maps and robustly restore the missing face regions even under large face pose and expression variations.

For methods targeting at single face [1], [47], they do not retrieve useful information from other frames but merely synthesize the missing regions for the current frame. Hence, they treat all the testing cases (shifting & static) the same way. It is obvious that they all failed to generate temporal consistent contents for the missing regions. Note that LaFin [47] and STN-GAN [65] also utilize face prior (i.e., landmarks) as their guidance. However, since the inpainting branch heavily depends on the landmark detection results, it will generate obvious artifacts when the predicted landmarks are incorrect (see Fig. 4).

C. Analysis of the Proposed Framework

In this section, we present experimental results to verify the design of our framework.

1) Effectiveness of UV-map Completion: We first carried out analysis on the effectiveness of our UV-map completion stage. We considered three variants, namely (a) single frame without UV maps as guidance, (b) single frame with UV maps as guidance, and (c) multi-frame without UV maps as guidance. Results are shown in Fig. 5 and Table II. Our full model achieved the most plausible results compared to these variant models. It is also observed that the performance improved considerably with UV maps as guidance especially under the static mask setting.

2) Effectiveness of FA Module: We also conducted ablation study to evaluate our FA module. For comparison, we considered three different baselines, namely (a) simply taking a single frame as input, (b) concatenating the target frame with its reference frames as input, and (c) fusing (concatenating) the features of all the frames in the latent space before decoder. The quantitative analysis evaluated on $\mathbf{U}_{out}$ are

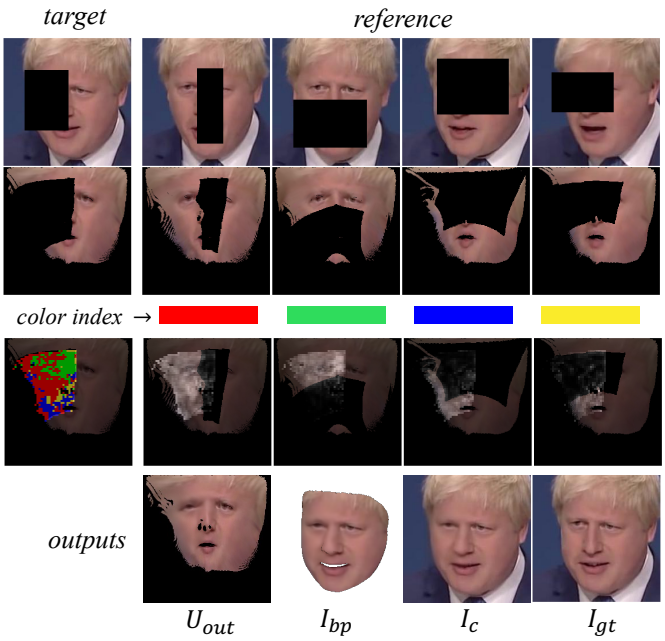


Fig. 7. Visualization of our Frame-wise Attention module. From top to bottom: input frame $\mathbf{I}_{in}$, input UV map $\mathbf{U}_{in}$, color index, attention map, and outputs. For the upper part, the first column is the *target* frame, the right four columns are *reference* frames.

listed in Table III. Our method achieved the best performance with the assistance of Frame-wise Attention. Referring to the qualitative results shown in Fig. 6, it is observed that our full model outperformed all the others in both detail generation and texture consistency, which also demonstrates the effectiveness of the FA module in retrieving proper correspondences for corrupted regions.

3) Visualization of Frame-wise Attention: To further investigate how does the FA module work, we present the visualization of the Frame-wise Attention in Fig. 7. We labeled each reference frame with a distinct color to visualize the attention map in a more intuitive way. For each *query* point from the embedded features in the target frame, we selected the most responsive *key* point (maximum attention value) from its pool of *key* candidates, and filled the attention map with the index color of the corresponding reference frame. In this example, we used the colors {*red*, *green*, *blue*, *yellow*} to denote the reference frames from left to right. The attention distribution is shown in the first column with the representative colors, while in the right four columns we display the response map of each reference frame. From the attention distribution, it is observed that the model learns to retrieve the matching features from the regions with higher reliability, *i.e.*, intact regions. With the FA module, our proposed MUC-Net can better exploit the reference features and generate more visually plausible content for the corrupted face.

4) Ablation Study on UV-map completion Stage: As mentioned in Sec. III-C, we take both flipped UV map $\hat{\mathbf{U}}_{in}$, the synthesized texture map $\mathbf{U}_t$, and the valid projection $\mathbf{U}_v$ as input. To further evaluate their contributions, we conducted ablation study on these components. Both quantitative results

TABLE IV
Quantitative results of different patch sizes on shifting masks.

Patch Size	$\ell_1 \downarrow$	PSNR ↑	SSIM ↑	VFID ↓
$s = 1$	0.0404	24.63	0.9720	0.2161
$s = 3$(ours)	**0.0394**	**24.80**	**0.9731**	**0.2007**
$s = 5$	0.0398	24.76	0.9727	0.2053
$s = 7$	0.0398	24.69	0.9727	0.2104

Fig. 8. Three examples of the averaged UV maps (per-pixel mean values of 5 consecutive frames). Small misalignment can be observed around the eye regions.

in Table III and qualitative results in Fig. 6 demonstrate their effectiveness in reconstructing the face textures by utilizing the symmetry prior ($\hat{\mathbf{U}}_{in}$) and 3D face model prior ($\mathbf{U}_t$). While $\mathbf{U}_v$ which indicates the valid face regions of the UV texture can help stabilize the training process and improve the overall performance.

5) Analysis on Patch Size used in Frame-wise Attention Module: Our method utilizes 3DMM face model as a bridge to transform the face textures from image space to UV space. Though the retrained face reconstruction network is capable of reconstructing proper face shapes for the corrupted input faces (refer to supplementary), it is possible that the predicted faces are slightly misaligned, which may result in small misalignment in the transformed UV maps. Fig. 8 shows the mean value of a bunch of inputs (target frame and its reference frames). We can see that there exists some small inconsistency especially around the eye regions. Therefore, for each *query* pixel, we propose to extract reference features in a local $s \times s$ window across all the reference frames. We further analyzed the effects of different patch sizes on shifting masks to observe how it affects the correspondence retrieval efficiency. Qualitative and quantitative results are shown in Fig. 9 and Table IV respectively. It is observed that adopting local windows instead of a single point can benefit the correspondence retrieval (the attention is more concentrated instead of scattered across the frames) and improve the overall performance. In our experiments, we adopted patch size $s = 3$ to achieve a balance between performance and efficiency.

6) Speed: We also estimated the processing speed of our method to assess its applicability. Our model achieved 19.3 fps with an NVIDIA GTX 2080Ti GPU card. Despite the primary goal of improving the inpainting quality for face videos, our method still achieves reasonable efficiency with a naïve implementation. Specifically, Resnet-50 as feature extractor occupies 3.3% of the time consumption, and the two main networks MUC-Net and FVR-Net take 45.4% in total,

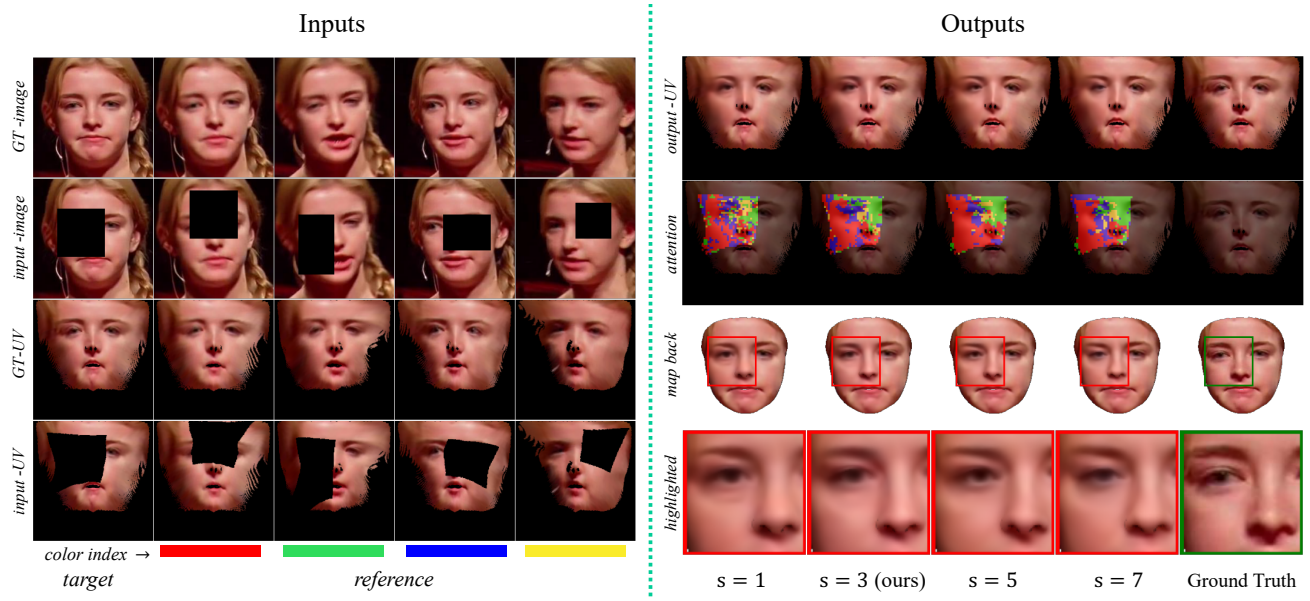


Fig. 9. Qualitative analysis on different patch size used in our Frame-wise Attention module. The left columns shows the masked input and ground truth of the *target* frame and *reference* frames, where from top to bottom: ground-truth image, masked image input, ground-truth UV map, masked UV input; The right side displays the results generated from different models with different patch sizes, where from top to bottom: output UV map, attention map, mapped back face and the highlighted regions.

TABLE V
Inference speed of video inpainting methods.

Method	3DGated [1]	STTN [86]	FuseFormer [90]	Ours
Speed (fps)	15.7	22.4	26.1	19.3

while the remaining 51.3% are for rendering process in UV mapping. We also compare our method with three recent video inpainting methods on inference speed. Results are shown in Table V. Please note that the rendering process in UV mapping accounts for 51.3% of the total time consumption. Nevertheless, our inference speed is still comparable to other baselines. In real applications, the rendering process can be further optimized and greatly accelerated.

D. User Study

We conducted a user study to further evaluate the visual quality of the inpainted videos. For comparison, we chose one image-based method LaFin [47] with landmarks as guidance, and two video-based methods – 3DGated [1], and STTN [86] with relatively higher performance. We sampled 16 videos from the test dataset, and tested on both static mask and shifting mask to evaluate the performance on these two cases. For each case, we sampled clips lasting 10 seconds from either rectangle mask or irregular mask (8 for each). The comparison is conducted in one-to-one manner with totally $3 \times 2 \times 16 = 96$ questions. For each question, the volunteers were given both the masked video and ground-truth video for reference, and were required to pick the better one from two inpainted videos (one baseline and ours). We collected responses from 20 volunteers and visualized the results in percentage (see Fig. 10). Our method gained most of the preference compared

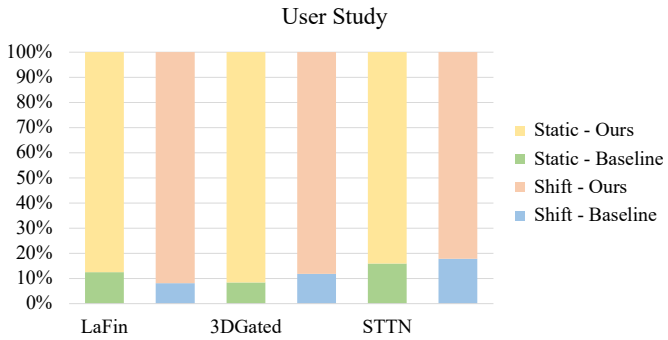


Fig. 10. User study results. We test on both static mask and shifting mask to evaluate the visual quality of our method. The test is conducted in one-to-one manner. For each baseline, left column is the result of static mask, and right column is for shifting mask. For each column, the upper stack shows the preference of our method, while the bottom stack shows the percentage for each baseline.

to other methods, which further demonstrates the effectiveness of our method.

E. Application

Face video inpainting usually serves as a recovering tool in many applications, such as video editing or restoration. It can be used to remove unwanted watermark / subtitles or objects that appear in face videos. An example is shown in Fig. 11 demonstrating the watermark removal application. Since our method is capable of both shifting and static masks with arbitrary shapes, it can benefit diverse face video editing tasks especially for those with large pose / expression variations (e.g., talk show).

(a) Input (b) Mask (red) (c) Ours (d) Ground truth

Fig. 11. Example of video watermark removal.

(a) Input (b) 3DMM (b) UV map (c) Ours (d) Ground truth

Fig. 12. Failure cases of our method.

F. Failure Case & Future Work

Since our method utilizes face model to explore the underlying 3D structure of the given corrupted faces, it is possible the predicted 3DMM is not perfectly fitted to the ground truth face, especially when the mask covers key clues for accurate alignment. As shown in Fig. 12, the eyes and nose are masked in the profile face, thus making it ambiguous for face reconstruction. The misaligned 3DMM (especially for nose region) results in noisy texture in the UV map and distorted nose in the final output. Currently, our second stage can help deal with small misalignment to refine the results. In our future work, we will try to improve the robustness of masked face reconstruction. Moreover, we will also extend this work to high-resolution face videos.

V. Conclusion

In this paper, we propose a novel approach to facilitate face video inpainting by exploring face texture completion in the UV space. The symmetry and aligned distribution of face textures in the UV space help to restore the masked regions with detailed face textures and structures. We design a Multi-reference UV-map Completion Network with a Frame-wise Attention module to enable efficient frame-wise correspondence retrieval from reference UV texture maps. Compared with existing state-of-the-art methods, our approach is capable of synthesizing more visually plausible results especially under large face pose and expression variations.

Acknowledgment

Guanying Chen is supported in part by NSFC with Grant No. 62293482 and 62202409, and the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001)

References

[1] Y.-L. Chang, Z. Y. Liu, K.-Y. Lee, and W. Hsu, "Free-form video inpainting with 3d gated convolution and temporal patchgan," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 9066–9075.

[2] J. Mathai, I. Masi, and W. AbdAlmageed, "Does generative face completion help face recognition?" in *International Conference on Biometrics (ICB)*. IEEE, 2019, pp. 1–8.

[3] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf, "Temporally coherent completion of dynamic video," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 1–11, 2016.

[4] A. Newson, A. Almansa, M. Fradet, Y. Gousseau, and P. Pérez, "Video inpainting of complex scenes," *Siam Journal on Imaging Sciences*, vol. 7, no. 4, pp. 1993–2019, 2014.

[5] M. Ebdelli, O. Le Meur, and C. Guillemot, "Video inpainting with short-term windows: Application to object removal and error concealment," *IEEE Transactions on Image Processing (TIP)*, vol. 24, no. 10, pp. 3034–3047, 2015.

[6] K. A. Patwardhan, G. Sapiro, and M. Bertalmio, "Video inpainting under constrained camera motion," *IEEE Transactions on Image Processing (TIP)*, vol. 16, no. 2, pp. 545–553, 2007.

[7] Y.-L. Chang, Z. Y. Liu, K.-Y. Lee, and W. Hsu, "Learnable gated temporal shift module for deep video inpainting," in *British Machine Vision Conference (BMVC)*, 2019.

[8] Y. Ding, C. Wang, H. Huang, J. Liu, J. Wang, and L. Wang, "Frame-recurrent video inpainting by robust optical flow inference," *arXiv preprint arXiv:1905.02882*, 2019.

[9] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Deep video inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5792–5801.

[10] S. Lee, S. W. Oh, D. Won, and S. J. Kim, "Copy-and-paste networks for deep video inpainting," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 4413–4421.

[11] A. Li, S. Zhao, X. Ma, M. Gong, J. Qi, R. Zhang, D. Tao, and R. Kotagiri, "Short-term and long-term context aggregation network for video inpainting," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 728–743.

[12] C. Wang, H. Huang, X. Han, and J. Wang, "Video inpainting by jointly learning temporal structure and spatial details," in *The AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 5232–5239.

[13] R. Xu, X. Li, B. Zhou, and C. C. Loy, "Deep flow-guided video inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3723–3732.

[14] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *ACM SIGGRAPH*, 1999, pp. 187–194.

[15] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *ACM SIGGRAPH*, 2000, pp. 417–424.

[16] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, "Filling-in by joint interpolation of vector fields and gray levels," *IEEE Transactions on Image Processing (TIP)*, vol. 10, no. 8, pp. 1200–1211, 2001.

[17] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, "Simultaneous structure and texture image inpainting," *IEEE Transactions on Image Processing (TIP)*, vol. 12, no. 8, pp. 882–889, 2003.

[18] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on Image Processing (TIP)*, vol. 13, no. 9, pp. 1200–1212, 2004.

[19] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Transactions on Graphics (TOG)*, vol. 28, no. 3, p. 24, 2009.

[20] A. Criminisi, P. Perez, and K. Toyama, "Object removal by exemplar-based inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2. IEEE, 2003, pp. II–II.

[21] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *ACM SIGGRAPH*, 2001, pp. 341–346.

[22] J. Hays and A. A. Efros, "Scene completion using millions of photographs," *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3, pp. 4–es, 2007.

[23] B.-W. Hwang and S.-W. Lee, "Reconstruction of partially damaged face images based on a morphable face model," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 25, no. 3, pp. 365–372, 2003.

[24] Z. Mo, J. P. Lewis, and U. Neumann, "Face inpainting with local linear representations." in *British Machine Vision Conference (BMVC)*, vol. 1, 2004, p. 2.

[25] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–14, 2017.

[26] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *IEEE Conference*

on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2536–2544.

[27] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6721–6729.

[28] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan, "Shift-net: Image inpainting via deep feature rearrangement," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 1–17.

[29] D. Wang, C. Xie, S. Liu, Z. Niu, and W. Zuo, "Image inpainting with edge-guided learnable bidirectional attention maps," *arXiv preprint arXiv:2104.12087*, 2021.

[30] C. Xie, S. Liu, C. Li, M.-M. Cheng, W. Zuo, X. Liu, S. Wen, and E. Ding, "Image inpainting with learnable bidirectional attention maps," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 8858–8867.

[31] H. Zhang, Z. Hu, C. Luo, W. Zuo, and M. Wang, "Semantic image inpainting with progressive generative networks," in *ACM International Conference on Multimedia*, 2018, pp. 1939–1947.

[32] M. Shao, W. Zhang, W. Zuo, and D. Meng, "Multi-scale generative adversarial inpainting network based on cross-layer attention transfer mechanism," *Knowledge-Based Systems*, vol. 196, p. 105778, 2020.

[33] D. Jia, N. Li, C. Li, S. Li, and W. Zuo, "Reflection removal of tongue image via total variation-based image inpainting," in *International Conference on E-Product E-Service and E-Entertainment*. IEEE, 2010, pp. 1–4.

[34] C. Wang, M. Shao, D. Meng, and W. Zuo, "Dual-pyramidal image inpainting with dynamic normalization," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[35] M. Ni, C. Wu, H. Huang, D. Jiang, W. Zuo, and N. Duan, "Nüwa-lip: Language guided image inpainting with defect-free vqgan," *arXiv preprint arXiv:2202.05009*, 2022.

[36] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "Edge-connect: Generative image inpainting with adversarial edge learning," *arXiv preprint arXiv:1901.00212*, 2019.

[37] Y. Song, C. Yang, Y. Shen, P. Wang, Q. Huang, and C.-C. J. Kuo, "Spg-net: Segmentation prediction and guidance network for image inpainting," *arXiv preprint arXiv:1805.03356*, 2018.

[38] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, "Structureflow: Image inpainting via structure-aware appearance flow," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 181–190.

[39] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, and J. Luo, "Foreground-aware image inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5840–5848.

[40] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Learning pyramid-context encoder network for high-quality image inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1486–1494.

[41] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4170–4179.

[42] M.-c. Sagong, Y.-g. Shin, S.-w. Kim, S. Park, and S.-j. Ko, "Pepsi: Fast image inpainting with parallel decoding network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 360–11 368.

[43] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5505–5514.

[44] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 85–100.

[45] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 4471–4480.

[46] A. Lahiri, A. Jain, D. Nadendla, and P. K. Biswas, "Improved techniques for gan based facial inpainting," *arXiv preprint arXiv:1810.08774*, 2018.

[47] Y. Yang, X. Guo, J. Ma, L. Ma, and H. Ling, "Lafin: Generative landmark guided face inpainting," *arXiv preprint arXiv:1911.11394*, 2019.

[48] X. Zhang, X. Wang, B. Kong, Y. Yin, Q. Song, S. Lyu, J. Lv, C. Shi, and X. Li, "Domain embedded multi-model generative adversarial networks for image-based face inpainting," *arXiv preprint arXiv:2002.02909*, 2020.

[49] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3911–3919.

[50] L. Song, J. Cao, L. Song, Y. Hu, and R. He, "Geometry-aware face completion and editing," in *The AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 2506–2513.

[51] Y. Zhao, W. Chen, J. Xing, X. Li, Z. Bessinger, F. Liu, W. Zuo, and R. Yang, "Identity preserving face completion for large ocular region occlusion," in *British Machine Vision Conference (BMVC)*, 2018.

[52] Y. Jo and J. Park, "Sc-fegan: Face editing generative adversarial network with user's sketch and color," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 1745–1753.

[53] T. Portenier, Q. Hu, A. Szabo, S. A. Bigdeli, P. Favaro, and M. Zwicker, "Faceshop: Deep sketch-based face image editing," *arXiv preprint arXiv:1804.08972*, 2018.

[54] X. Li, G. Hu, J. Zhu, W. Zuo, M. Wang, and L. Zhang, "Learning symmetry consistent deep cnns for face completion," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 7641–7655, 2020.

[55] X. Li, M. Liu, Y. Ye, W. Zuo, L. Lin, and R. Yang, "Learning warped guidance for blind face restoration," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 272–289.

[56] X. Li, W. Li, D. Ren, H. Zhang, M. Wang, and W. Zuo, "Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2706–2715.

[57] X. Li, C. Chen, S. Zhou, X. Lin, W. Zuo, and L. Zhang, "Blind face restoration via deep multi-scale component dictionaries," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 399–415.

[58] C. Chen, X. Li, L. Yang, X. Lin, L. Zhang, and K. Wong, "Progressive semantic-aware style transformation for blind face restoration," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11 896–11 905.

[59] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9168–9178.

[60] K. A. Patwardhan, G. Sapiro, and M. Bertalmio, "Video inpainting of occluding and occluded objects," in *IEEE International Conference on Image Processing (ICIP)*, vol. 2. IEEE, 2005, pp. II–69.

[61] Y. Wexler, E. Shechtman, and M. Irani, "Space-time completion of video," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 29, no. 3, pp. 463–476, 2007.

[62] C. Gao, A. Saraf, J.-B. Huang, and J. Kopf, "Flow-edge guided video completion," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 713–729.

[63] X. Zou, L. Yang, D. Liu, and Y. J. Lee, "Progressive temporal feature alignment network for video inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 448–16 457.

[64] S. W. Oh, S. Lee, J.-Y. Lee, and S. J. Kim, "Onion-peel networks for deep video completion," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 4403–4412.

[65] Y. Wu, V. Singh, and A. Kapoor, "From image to video face inpainting: Spatial-temporal nested gan (stn-gan) for usability recovery," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 2396–2405.

[66] X. Wu, R.-L. Li, F.-L. Zhang, J.-C. Liu, J. Wang, A. Shamir, and S.-M. Hu, "Deep portrait image completion and extrapolation," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 2344–2355, 2019.

[67] X. Yuan and I. K. Park, "Face de-occlusion using 3d morphable model and generative adversarial network," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 10 062–10 071.

[68] X. Zeng, X. Peng, and Y. Qiao, "Df2net: A dense-fine-finer network for detailed 3d face reconstruction," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 2315–2324.

[69] B. Chaudhuri, N. Vesdapunt, L. Shapiro, and B. Wang, "Personalized face modeling for improved face reconstruction and motion retargeting," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 142–160.

[70] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[71] B. Egger, W. A. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani *et al.*, "3d morphable face models—past, present, and future," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 5, pp. 1–38, 2020.

[72] E. Sariyanidi, C. J. Zampella, R. T. Schultz, and B. Tunc, "Inequality-constrained and robust 3d face model fitting," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 433–449.

[73] J. Lin, Y. Yuan, T. Shao, and K. Zhou, "Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5891–5900.

[74] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou, "Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7093–7102.

[75] B. Gecer, J. Deng, and S. Zafeiriou, "Ostec: One-shot texture completion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 7628–7638.

[76] M. Cao, H. Huang, H. Wang, X. Wang, L. Shen, S. Wang, L. Bao, Z. Li, and J. Luo, "Task-agnostic temporally consistent facial video editing," *arXiv preprint arXiv:2007.01466*, 2020.

[77] T. Nguyen, A. T. Tran, and M. Hoai, "Lipstick ain't enough: Beyond color matching for in-the-wild makeup transfer," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 305–13 314.

[78] S. Xu, J. Yang, D. Chen, F. Wen, Y. Deng, Y. Jia, and X. Tong, "Deep 3d portrait from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7710–7720.

[79] X. Hu, W. Ren, J. LaMaster, X. Cao, X. Li, Z. Li, B. Menze, and W. Liu, "Face super-resolution guided by 3d facial priors," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 763–780.

[80] W. Ren, J. Yang, S. Deng, D. Wipf, X. Cao, and X. Tong, "Face video deblurring using 3d facial priors," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 9388–9397.

[81] W.-S. Lee, P. Kalra, and N. Magnenat-Thalmann, "Model based face reconstruction for animation," in *International Conference on Multimedia Modeling (MMM)*, vol. 97. Citeseer, 1997, pp. 323–338.

[82] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[83] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2009, pp. 296–301.

[84] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2013.

[85] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803.

[86] Y. Zeng, J. Fu, and H. Chao, "Learning joint spatial-temporal transformations for video inpainting," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 528–543.

[87] 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 (TIP)*, vol. 13, no. 4, pp. 600–612, 2004.

[88] S. Zhao, J. Cui, Y. Sheng, Y. Dong, X. Liang, E. I. Chang, and Y. Xu, "Large scale image completion via co-modulated generative adversarial networks," in *International Conference on Learning Representations (ICLR)*, 2021.

[89] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mat: Mask-aware transformer for large hole image inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 758–10 768.

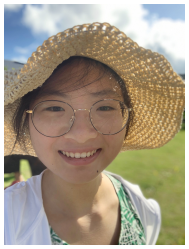
[90] R. Liu, H. Deng, Y. Huang, X. Shi, L. Lu, W. Sun, X. Wang, J. Dai, and H. Li, "Fuseformer: Fusing fine-grained information in transformers for video inpainting," in *International Conference on Computer Vision (ICCV)*, 2021, pp. 14 040–14 049.

[91] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[92] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *IEEE International Conference on Computer Vision (ICCV) Workshops*, 2015, pp. 50–58.

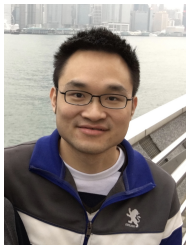
[93] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[94] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision (ICCV)*, 2017, pp. 1021–1030.



Wenqi Yang is currently a Ph.D. candidate at the Department of Computer Science, The University of Hong Kong, supervised by Dr. Kenneth K.Y. Wong. She received her B.Eng. degree from Zhejiang University in 2019. Her research interests are Computer Vision and Deep Learning.

Zhenfang Chen is a researcher at MIT-IBM Watson AI Lab in Cambridge, MA, USA. He received his Ph.D. degree from the Department of Computer Science at The University of Hong Kong, where he was a member of the Computer Vision Lab, advised by Prof. Kenneth K.Y. Wong. Prior to that, he got his B.Sc. from Sun Yat-sen University in 2016. His interests are centered around machine learning and its applications to computer vision and natural language processing.



Chaofeng Chen is currently a postdoctoral research fellow at Nanyang Technological University. He received his Ph.D. degree from Department of Computer Science at the University of Hong Kong in 2021. Prior to that, he obtained his B.Eng. from Huazhong University of Science and Technology. His interests are centered around Computer Vision and Deep Learning.

Guanying Chen received his B.Eng. degree from Sun Yat-sen University in 2016, and the Ph.D. degree in the Department of Computer Science from the University of Hong Kong in 2021. He is currently a Research Assistant Professor at The Chinese University of Hong Kong, Shenzhen. His research interests are learning based methods for computer vision.

Kwan-Yee K. Wong (Senior Member, IEEE) received the B.Eng. degree (Hons.) in computer engineering from The Chinese University of Hong Kong in 1998, and the M.Phil. and Ph.D. degrees in computer vision (information engineering) from the University of Cambridge in 2000 and 2001, respectively. Since 2001, he has been with the Department of Computer Science at The University of Hong Kong, where he is currently an Associate Professor. His research interests are in computer vision and machine intelligence. He is currently an editorial board member of International Journal of Computer Vision (IJCV).