NON-DETERMINISTIC FACE MASK REMOVAL BASED ON 3D PRIORS

Xiangnan YIN and Liming CHEN

Ecole Centrale de Lyon

ABSTRACT

This paper presents a novel image inpainting framework for face mask removal. Although current methods have demonstrated their impressive ability in recovering damaged face images, they suffer from two main problems: the dependence on manually labeled missing regions and the deterministic result corresponding to each input. The proposed approach tackles these problems by integrating a multi-task 3D face reconstruction module with a face inpainting module. Given a masked face image, the former predicts a 3DMM-based reconstructed face together with a binary occlusion map, providing dense geometrical and textural priors that greatly facilitate the inpainting task of the latter. By gradually controlling the 3D shape parameters, our method generates high-quality dynamic inpainting results with different expressions and mouth movements. Qualitative and quantitative experiments verify the effectiveness of the proposed method.

Index Terms— mask removal, face inpainting, 3DMM

1. INTRODUCTION

Wearing face masks in public has become an essential hygiene practice to control the spread of COVID-19, posing new challenges for face-related computer vision tasks. Computers need to accomplish face recognition, expression recognition, landmark detection, etc., using minimal exposed facial textures. Although many recent studies focus on the masked scenario, most are task-specific and not universally applicable. In comparison, directly restoring mask-occluded face texture promises to be a one-stop solution to the problem. To this end, we need to tackle two sub-tasks: 1) detecting the occluded region and 2) recovering the face textures, corresponding to image segmentation and face image inpainting, respectively.

Thanks to the revolutionary emergence of deeplearning, data-driven approaches have dominated computer vision with great success. However, this also led to the reliance on high-quality training data. Regarding mask segmentation specifically, large, diverse, and manually annotated mask datasets are in strong demand due to the targets' varying shapes, orientations, and textures. Some methods synthesize training data by overlaying masks on ordinary face images. As an interim solution, it would be sensible to adopt a similar strategy until the real large face mask parsing datasets become available.

Assuming that missing regions share similar content to the visible regions, early image inpainting methods fill the holes by iteratively searching nearest neighbor textures from the background [1]. However, such copy-and-paste methods only consider internal information within the image, making them only capable of recovering tiny, smooth textures and not dealing with semantic-level deficiencies such as masked noses and mouths. On the other hand, data-driven approaches learn the data distribution from large datasets, allowing them to restore the semantic-level image patterns. SSDA [2] introduced the deep autoencoder to image inpainting for the first time, and Context Encoder [3] pioneered the autoencoder-discriminator training paradigm. [4, 5] exploits feature masking to deal with freeform missing regions. Also, different attention modules [6, 7, 8] have been proposed to break through the limited receptive field of the convolution kernel and thus explicitly model long-distance dependencies. Despite efforts to improve inpainting quality, most approaches produce only deterministic results, which is not reasonable, as the masked areas can have multiple filling options.

This paper proposes a novel 3D reconstruction-guided method for removing masks from face images in the wild. The model comprises a multi-task mask-robust 3D face reconstruction module and a face inpainting module. The former predicts both the 3D Morphable Model (3DMM) [9] parameters and the binary occlusion map of the masked face, and the latter recovers the missing facial texture conditioned by the rendered 3D prior. By changing the 3DMM parameters, we can control the shape and expression of the recovered face both accurately and smoothly.

2. RELATED WORKS

The closest work to ours is that of Din *et al.* [10], where we both focus on the problem of face mask removal and divide it into mask segmentation and face painting. Our method surpasses theirs in two aspects: 1) We labeled more mask templates (900 vs. 50) to train the mask segmentation task. 2) Our inpainting results are diverse and highly controllable.

Some variational autoencoder (VAE)-based methods can also produce non-deterministic outputs [11, 8] by sampling latent codes from predicted distributions. Although the stochastic nature of the VAE brings about varied results, diversity is never guaranteed: the targets are still fixed, lead-

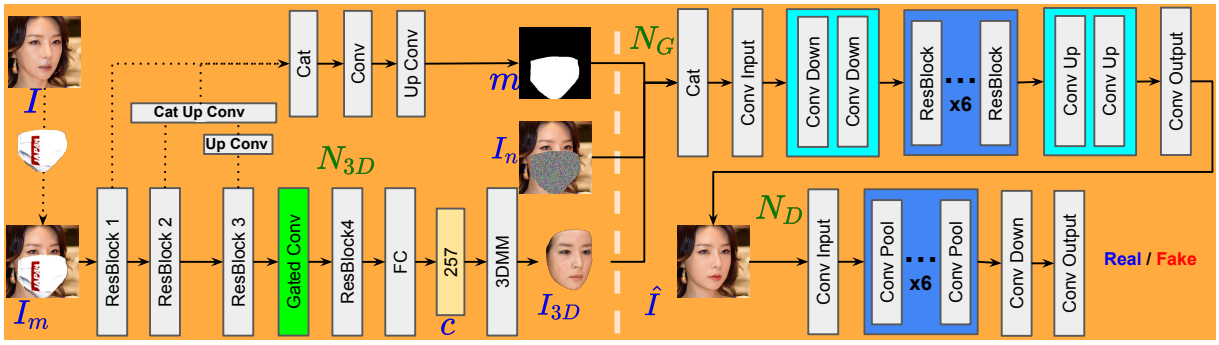

Fig. 1. We first synthesize the training data I_m by adding masks to ordinary face images I. Next, we use a multi-task model N_{3D} to predict the mask silhouette m and the 3D reconstructed face I_{3D} of the input. Finally, we synthesize the non-occluded face $\hat{I}$ based on the noised face I_n, m and I_{3D}. A VGG-shaped discriminator N_D is leveraged to distinguish $\hat{I}$ from the real.

ing to 1) sharp latent distributions and 2) robust decoder to the latent codes' variations, degrading the framework into a common autoencoder. Furthermore, the diversity introduced by random sampling is neither controllable nor smooth. Although some other methods conditioned on sketches [12], facial landmarks [13], or segmentation maps [14] do yield editable results, the sparsity and instability of such conditions lead to poor controlling accuracy.

3. APPROACH

Since collecting large amounts of paired with/without the mask face images is infeasible, we train on synthetic data pairs generated by overlaying masks on ordinary face images, as shown on the leftmost of Figure 1. The proposed model is composed of a multi-task 3D face reconstruction-mask segmentation module N_{3D} and a face inpainting module N_G, corresponding to the left and right halves of Figure 1, respectively. Given a masked face image I_m, N_{3D} predicts its 1) corresponding 3DMM parameters c, from which a 3D face I_{3D} could be reconstructed and rendered, and 2) the occlusion mask m, indicating the mask silhouette. We then replace the mask texture with random noises based on m and get I_n. Finally, N_G predicts the mask-free face $\hat{I}$ conditioned by I_n, m, and I_{3D}. We also employ a discriminator N_D to increase the realism of the generated images. The Following presents each module in detail.

3.1. Segmentation-Reconstruction Module

N_{3D} takes ResNet50 [15] as its backbone and fulfills 3D face reconstruction and mask segmentation. Intuitively, the neural network captures global shape patterns in the bottom layers and detailed texture patterns in the top layers, while the masks usually occupy a large area of the face with relatively simple textures, so we perform mask segmentation using features from the first three residual blocks. The segmentation task is solely guided by the Binary Cross-Entropy (BCE) loss:

$$L_{bce} = -\frac{1}{WH}\sum(m\odot\log(\hat{m})+(1-m)\odot\log(1-\hat{m})), \quad (1)$$

where $\hat{m}$ and m denote the predicted and the ground truth binary mask, W and H denote the spatial dimension of the binary map. The "online hard example mining" (OHEM) technique is also utilized to make the training more efficient.

To concentrate the model on visible textures while predicting the 3DMM parameters, we also integrate a gated convolution [5] layer before the last residual block of ResNet50, predicting a dynamic feature mask for each channel. The 3D reconstruction branch outputs a vector $\hat{c} \in \mathbb{R}^{237}$, containing the face's shape, pose, texture, and illumination parameters. To accelerate the training, we use 3D coefficients predicted from the original unmasked face images by the pre-trained model of [16] as the ground truth of c. Following losses jointly guide the 3D reconstruction task:

The most direct term is the coefficient loss.

$$L_{coef} = \frac{1}{N}\|\hat{c} - c\|_1, \quad (2)$$

where $\hat{c}$ and c are the predicted and the ground truth 3D coefficients, N denote the dimension of c.

However, the coefficient level loss treats the discrepancy in all dimensions equally, which is unreasonable, as some dimensions affect the reconstruction results much more than others (e.g., poses v.s. illuminations). Hence we introduce the photo loss, which constraints the training at the image level.

$$L_{photo} = \frac{1}{\sum M}\|I_{3D}\odot M - I\odot M\|_2, \quad (3)$$

where I_{3D} denotes the rendered reconstruction result, I denotes the original face image, and M denotes the binary face region map (provided by the training dataset).

As with most face reconstruction methods, we apply identity loss for better capturing the face identity.

$$L_{id} = 1 - \frac{\mathcal{F}(I_{3D})\mathcal{F}(I)}{\|\mathcal{F}(I_{3D})\|\|\mathcal{F}(I)\|}, \quad (4)$$

where $\mathcal{F}(\cdot)$ denotes the feature extraction operation via a pre-trained Arcface[17] model.

Finally, we leverage landmark loss as [16] to loosely constrain the shape and pose of the reconstructed face.

$$L_{lm} = \frac{1}{n_{pt}} \sum_{i=1}^{n_{pt}} \omega_i \|\hat{q}_i - q_i\|^2, \quad (5)$$

where $\hat{q}_i$ and q_i represent the 68 ($n_{pt} = 68$) facial landmarks indexed from the predicted and the ground truth (reconstructed from c in Equation 2) 3D faces, respectively. ω_i is the weight corresponding to the ith landmark, set to 20 for the nose and inner-mouth points and 1 for others.

The overall loss function is formulated as:

$$L_{3D} = L_{bce} + L_{coef} + L_{photo} + \lambda id L_{id} + \lambda_{lm} L_{lm}, \quad (6)$$

where $\lambda_{id} = 0.1$ and $\lambda_{lm} = 0.001$.

3.2. Inpainting Module

The inpainting module consists of a generator N_G with stacked residual blocks and a discriminator N_D with the VGG structure. As shown in Figure 1, N_G concatenates the mask parsing map m, the 3DMM-based face I_{3D}, and the noised image I_n as input and outputs $\hat{I}$, which recovers the original mask-free face image I. Further, $\hat{I}$ and I are fed into N_D to obtain their probabilities of being real data. We utilize the following losses to train the model:

Pixel-wise loss,

$$L_{pix} = \frac{1}{HWC} \|\hat{I} - I\|_1, \quad (7)$$

where H, W, C are the height, width and channels of I.

Identity loss L_{id}, formulated the same as Equation 4, except replacing I_{3D} therein with $\hat{I}$.

Total variation loss [18],

$$L_{tv} = \frac{1}{HWC} (\|\nabla_x \hat{I}\|^2 + \|\nabla_y \hat{I}\|^2), \quad (8)$$

where $\nabla_{\_}$ denotes the directional gradient.

Adversarial loss,

$$L_{adv} = -\mathbb{E}_{\hat{I}}[\log D(\hat{I}_i)], \quad (9)$$

where $D(\cdot)$ denotes the mapping function of N_D; the larger its value, the more its input tends to be real.

The full loss of N_G is summarized as:

$$L_G = \lambda_{pix} L_{pix} + \lambda_{id} L_{id} + \lambda_{tv} L_{tv} + \lambda_{adv} L_{adv}, \quad (10)$$

where $\lambda_{pix} = 10$, $\lambda_{id} = 0.1$, $\lambda_{tv} = 0.1$, $\lambda_{adv} = 0.01$.

Discriminator loss, the loss of N_D follows the implementation of [19], which is composed of an ordinary BCE loss and a zero-centered gradient penalty for real images,

$$L_D = \mathbb{E}_I[\log(D(I))] + \mathbb{E}_{\hat{I}}[\log(1 - D(\hat{I}))] + \mathbb{E}_I[\nabla_I^2 D(I)] \quad (11)$$

Methods	Hong *et al.*[20]	ELFW [21]	Din *et al.*[10]	Anwar *et al.*[22]	**Ours**
Shapes	14	12	50	20	**900**
Textures	–	–	–	27	**800**

Table 1. The mask diversity of different methods or datasets.

Fig. 2. 3D face reconstruction ability for masked faces.

4. EXPERIMENTS

We first present our implementation details. Then, we qualitatively compare our method's 3D face reconstruction, mask removal, and face editing abilities with state-of-the-art. Finally, we quantitatively compare the face restoration ability of different methods at pixel and perceptual levels.

4.1. Training Details

Most mask-related approaches synthesize masked/unmasked training pairs by overlaying mask templates on face images of existing face datasets. However, as Table 1 shows, the mask templates used by previous methods are pretty limited; only a few tens of variations are far from sufficient to train a robust model. Therefore, we 1) manually keyed out 900 masks from the masked face images and 2) collected 800 texture patches to replace the textures of the original masks[1].

The mask templates are then combined with CelebAMask-HQ [23] and FFHQ [24] to generate data pairs on the fly as training goes (1000 images from FFHQ are left out for testing). We train 500,000 steps for N_{3D} and 200,000 steps for N_G-N_D, both with a batch size of 8 and an initial learning rate of $1e^{-4}$. For each module, the learning rate drops to $1e^{-5}$ when the training reaches its midpoint. We use Adam with betas set to $[0.9, 0.999]$ to optimize the two modules. It takes about 60 hours to train N_{3D} and 40 hours to train N_G-N_D on two Nvidia GTX 1080 GPUs.

4.2. Qualitative results

Accurately reconstructing the 3D face from masked faces is the prerequisite for the success of the subsequent inpainting module. Therefore, we first compare our method with the SOTA 3D reconstruction method of Deng *et al.* As shown

[1]Images are downloaded from Google and labeled using Apple Pencil.

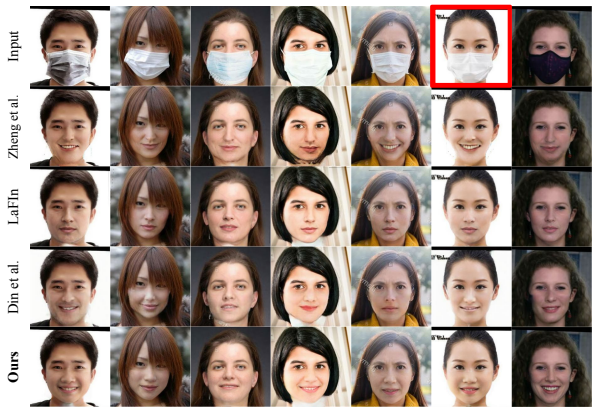

Fig. 3. Comparison of mask removal ability, the inputs are aligned according to the methods' settings, and the outputs are remapped to the original images for a consistent view.

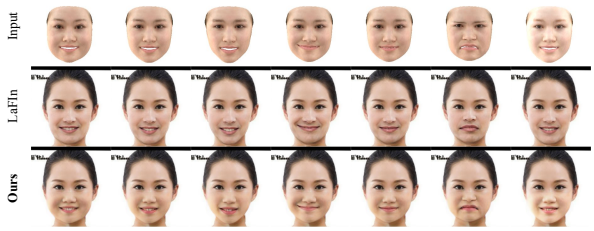

Fig. 4. Comparison of face editing ability.

in Figure 2, the method of Deng *et al.* is stronly influenced by the mask, resulting in deviations in texture and poses. In comparison, our method is robust to face masks thanks to the synthetic masked face images for training.

As Section 2 mentions, the method closest to ours is that of Din *et al.* [10]. Unfortunately, they do not released their code; therefore, we use the images from their paper for a more convincing comparison. The other two methods we compare are LaFIn [13] and Zhang *et al.*, which can generate diverse inpainting results. We provide those methods with mask regions detected by N_{3D}. As Figure 3 shows, our approach significantly outperforms Zhang *et al.* and Din *et al.*. The random sampling in the hidden space leads to apparent artifacts in the results of Zhang *et al.*. Without a shape prior, the method of Din *et al.* may generate distorted faces (row 4 column 2); In addition, the poor accuracy of their mask segmentation module results in residual mask edges on the face (row 4, columns 6 and 7). Our results are comparable with LaFIn, however, the latter requires additional binary mask maps.

We further compare the face editing ability of our model with LaFIn, the landmark-guided face inpainting method. This time we provide LaFIn with 68 facial landmarks extracted from our predicted 3D face model. Figure 4 shows the results guided by different 3D priors(for the face in the red box in Figure 3). The first six columns are conditioned by different shapes and the last column is conditioned by a brighter skin. As can be seen, with the guidance of our landmarks, LaFIn can generate diverse inpainting results. Nevertheless, due to the sparsity of the landmark, the generated faces do not precisely comply with the 3D face shapes; in addition, LaFIn cannot change the skin color as we do in the last column.

Methods	$L_1 \downarrow$	PSNR ↑	SSIM ↑	Cos ID ↑	FID ↓
Zheng *et al.*	0.020	24.709	0.893	0.506	12.808
LaFIn	0.018	25.569	0.897	0.543	12.465
Din *et al.*	–	26.19	0.864	–	**3.548**
Ours	**0.014**	**27.230**	**0.912**	**0.654**	9.744

Table 2. Quantitative comparison.

4.3. Quantitative results

We synthesized 1000 masked face images on the test set, and then used LaFIn and Zheng *et al.*'s method to recover the unmasked faces (with externally provided binary mask maps). The face restoration ability is evaluated by: L1 Loss, PSNR score, SSIM score, FID score, and the cosine similarity of the identity features extracted by [17]. For Din *et al.*'s method, since their code is not publicly available, we adopt the data from their paper. Results are shown in Table 2. As can be seen, our method outperforms others in all metrics except for its higher FID score than Din *et al.*; however, it should be noted that the generation quality of Din *et al.* is not as good as its FID score reflects.

5. CONCLUSION

This paper proposes a novel framework for removing the mask from the face image. First, we manually labeled a large, high-quality dataset of face masks for synthesizing training pairs. Next, we trained a mask-robust multi-task module for reconstructing 3D faces and detecting the mask region of face images. Finally, we proposed a 3D reconstruction guided face inpainting module to generate non-deterministic and highly-controllable results. The proposed method outperforms the state-of-the-art qualitatively and quantitatively.

6. REFERENCES

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 24, 2009.

[2] Junyuan Xie, Linli Xu, and Enhong Chen, "Image denoising and inpainting with deep neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[3] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros, "Context encoders:

Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[4] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 85–100.

[5] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4471–4480.

[6] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5505–5514.

[7] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*. PMLR, 2019, pp. 7354–7363.

[8] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai, "Pluralistic image completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1438–1447.

[9] Volker Blanz and Thomas Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194.

[10] Nizam Ud Din, Kamran Javed, Seho Bae, and Juneho Yi, "A novel gan-based network for unmasking of masked face," *IEEE Access*, vol. 8, pp. 44276–44287, 2020.

[11] Kihyuk Sohn, Honglak Lee, and Xinchen Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

[12] Youngjoo Jo and Jongyoul Park, "Sc-fegan: Face editing generative adversarial network with user's sketch and color," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1745–1753.

[13] Yang Yang, Xiaojie Guo, Jiayi Ma, Lin Ma, and Haibin Ling, "Lafin: Generative landmark guided face inpainting," *arXiv preprint arXiv:1911.11394*, 2019.

[14] Li Yu, Dequan Zhu, and Jian He, "Semantic segmentation guided face inpainting based on sn-patchgan," in *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2020, pp. 110–115.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[16] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[17] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[18] Aravindh Mahendran and Andrea Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5188–5196.

[19] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8188–8197.

[20] Je Hyeong Hong, Hanjo Kim, Minsoo Kim, Gi Pyo Nam, Junghyun Cho, Hyeong-Seok Ko, and Ig-Jae Kim, "A 3d model-based approach for fitting masks to faces in the wild," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 235–239.

[21] Rafael Redondo and Jaume Gibert, "Extended labeled faces in-the-wild (elfw): Augmenting classes for face segmentation," *arXiv preprint arXiv:2006.13980*, 2020.

[22] Aqeel Anwar and Arijit Raychowdhury, "Masked face recognition for secure authentication," *arXiv preprint arXiv:2008.11104*, 2020.

[23] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5549–5558.

[24] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.