HDTR-Net: A Real-Time High-Definition Teeth Restoration Network for Arbitrary Talking Face Generation Methods

Yongyuan Li[1,2], Xiuyuan Qin[3], Chao Liang[4], and Mingqiang Wei[1,2(✉)]

[1] Nanjing University of Aeronautics and Astronautics, Nanjing, China
`mqwei@nuaa.edu.cn`
[2] Shenzhen Research Institute, Nanjing University of Aeronautics and Astronautics, Shenzhen, China
[3] Soochow University, Suzhou, China
[4] Nanjing University of Science and Technology, Nanjing, China

Abstract. Talking Face Generation (TFG) aims to reconstruct facial movements to achieve high natural lip movements from audio and facial features that are under potential connections. Existing TFG methods have made significant advancements to produce natural and realistic images. However, most work rarely takes visual quality into consideration. It is challenging to ensure lip synchronization while avoiding visual quality degradation in cross-modal generation methods. To address this issue, we propose a universal High-Definition Teeth Restoration Network, dubbed HDTR-Net, for arbitrary TFG methods. HDTR-Net can enhance teeth regions at an extremely fast speed while maintaining synchronization, and temporal consistency. In particular, we propose a Fine-Grained Feature Fusion (FGFF) module to effectively capture fine texture feature information around teeth and surrounding regions, and use these features to fine-grain the feature map to enhance the clarity of teeth. Extensive experiments show that our method can be adapted to arbitrary TFG methods without suffering from lip synchronization and frame coherence. Another advantage of HDTR-Net is its real-time generation ability. Also under the condition of high-definition restoration of talking face video synthesis, its inference speed is 300% faster than the current state-of-the-art face restoration based on super-resolution. Our code and trained models are released at https://github.com/yylgoodlucky/HDTR.

Keywords: High-Definition Teeth Restoration Network · Talking Face Generation · Teeth Restoration · Visual Quality

1 Introduction

Talking Face Generation (TFG) plays an important role in the audio-visual field [26], as it enables the integration of visual and auditory information to enhance the understanding and perception of information for humans.

For TFG, a clear and realistic mouth in the generated image could provide a richer audio-visual experience and thus help the user to understand the semantics better. Traditional TFG methods warp the source image with the help of prior knowledge (e.g., audio), and many of them result in inaccurate output. Thanks to the successful application of deep neural networks, especially Generative Adversarial Networks (GAN) [17] and Convolution Neural Networks (CNN), TFG has made significant progress. Some efforts [22,24,27,29,39,41,44] try to produce natural and realistic talking faces by extracting driving features from speech signals and then integrate them into face animation. Existing TFG methods focus on producing natural realistic and high synchronization mouth shapes. However, it is still challenging to enhance the teeth clarity while ensuring the natural synchronization of mouth shapes.

Analyzing from the data perspective, the low-resolution of images in existing TFG datasets limits the ability of cutting-edge models to generate high-resolution mouth shapes. Analyzing from the network perspective, Obamanet et al. [25] propose a Teeth Proxy to obtain the high-frequency components of the teeth from the candidate frames to improve the clarity of the upper and lower teeth. However, these methods require a rigid selection of the candidate frames and do not optimize the clarity of the surrounding areas of the teeth. Some other efforts [22,42] use additional reference frames to compensate mouth shapes and motion to guide the network for accurate modeling of head posture and mouth synchronization, which produce excellent lip synchronization, but the model still falls short in terms of clarity in predicting teeth and their surrounding regions. We argue that prior knowledge is insufficient to provide and restore fine-grained features about the teeth and their surrounding regions.

Most recent efforts [5,18,31,32] recover high-frequency details from blurred and degraded face images while maintaining the original face features, but it still has gaps in terms of frame coherence and speed. In face restoration, the process is applied to each frame individually, which can result in noticeable pixel discontinuities in the synthesized video. This occurs because face restoration primarily focuses on recovering high-frequency information within each frame, without considering the continuity between frames. As a result, there may be visible inconsistencies or discrepancies in the appearance of consecutive frames, leading to a lack of smoothness or coherence in the synthesized video.

In addition, processing each frame takes a lot of time even on low-resolution images, which limits the practical applicability of face restoration in real-time or time-sensitive scenarios.

We propose a universal real-time High-Definition Teeth Restoration Network (HDTR-Net). HDTR-Net can be applied to arbitrary Talking Face Generation methods for generating quality results (see Fig. 1). Two components are involved in the proposed method: a Fine-Grained Feature Fusion (FGFF) module and a Decoder module. We deliberately design the FGFF module to merge features responsible for extracting the image texture details. The branch below the FGFF module is utilized to leverage the reference image as guidance, enabling the model to effectively restore high-frequency details. In addition, it can be used as a prompt for pixel continuity in the inference stage. The Decoder module
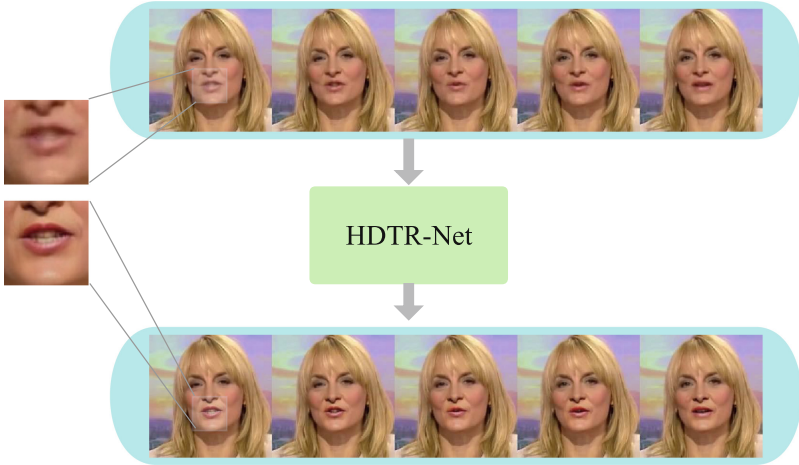
Fig. 1. method effectively enhances the clarity of teeth and surroundings generated by arbitrary talking face videos.

restores the high-frequency details from the extracted fine-grained feature. The main contributions of our work are three-fold:

- We propose a High-Definition Teeth Restoration Network, dubbed HDTR-Net, to enhance the clarity of teeth regions while maintaining texture details. HDTR-Net can be applied to arbitrary talking face generation.
- We propose a Fine-Grained Feature Fusion module, which has the ability to extract fine-grained features effectively.
- HDTR-Net exhibits exceptional speed in terms of repairing capability. For example, our inference speed is 300% faster than the super-resolution based image restoration methods.

2   Related Work

2.1   Talking Face Generation

Talking Face Generation aims to synthesize a sequence of talking face frames according to a sequence of driving audio or text, which is a multi-modal learning method for mapping acoustic features to real facial motions. In recent years, [15,25] have learned implicit mapping functions from audio to corresponding landmarks of the mouth to implement talking face synthesis, but their work is only specific to a particular speaker and audio. It is not sufficient to extend to arbitrary speakers and audio, and the synthesized video has low-quality clarity. [10] proposes a deblurring module, which uses the idea of early super-resolution [12] to transfer the facial feature map from the input image to the generated output using a skip connection to avoid producing a blurred image. Subsequently, [22,27,28,44] train on the large-scale public dataset (e.g., LRS2 [1], LRW [6])

in an end-to-end manner and produce effective results, being able to generalize to arbitrary speakers and audio, but the mouth region is still of low resolution in the synthesized video, resulting in a worse visual experience for the viewer. [4,8,43] firstly allow the speech to predict the face landmarks, and then recover the real face image. Predicting face landmarks couple speech features and mouth motion features obtains better lip synchronization, but the two-stage training method inherently loses information. In order to synthesize more natural and realistic head motions, the reference sequence images are fed into the network to reduce the head pose and mouth motions [22,30,41,42]. Respectively, [42] and [41] decoupled the visual representation space using contrastive learning [20] and generative adversarial methods [17], which produce significant improvements in lip synchronization. To obtain high-fidelity talking face videos, [38] proposes a repair network with an adaptive affine transformation module [37] to achieve clearly synthetic videos with multi-stage training, but the inference speed is slow and the fidelity of the mouth shape depends on the clarity of the input image.

2.2   Face Restoration

Based on the general face hallucination [2,5,9,34,35], most methods utilize a geometric prior and a reference prior together to improve performance. However, geometric priors are estimated from low-quality input images, and such geometric priors cannot provide detailed information about the image. Reference prior [7,18,19] relies on images of the same identity, and its ability to recover high-frequency details is reduced for particular image feature with rich appearance. In particular, [14] estimates face landmarks before restoring a face and estimating facial pose, while for quite minor facial accurate estimation is difficult. [45] proposes a unified framework for multi-resolution and dense correspondence field estimation of faces to recover texture details. Recently, deep learning-based models have advanced significantly in image processing tasks and are at present driving the state-of-the-art in face restoration. [16,23] have better reconstruction performance with a generative approach. [36] uses the discriminative generative network to ultra-resolve images by aligning miniature low-resolution face images. [40] uses convolution to extract features from blurred images to reconstruct high-definition face images, but their restored faces generate unfaithful outcomes. These methods are essentially based on restoration for single images, with the benefit of being able to recover a scaled-up, high-resolution image, however, the temporal coherence present in the video is not taken into account.

3   Method

We propose a novel real-time teeth restoration network called HDTR-Net that can be adaptive to arbitrary TFG methods. The structural details of HDTR-Net are shown in Fig. 2. We first review the structure of the model, which contains two parallel Fine-Grained Feature Fusion (FGFF) modules and a Decoding module. Channel Fusion (CF) and HourGlass are included in FGFF, which focus on extracting fine-grained features.
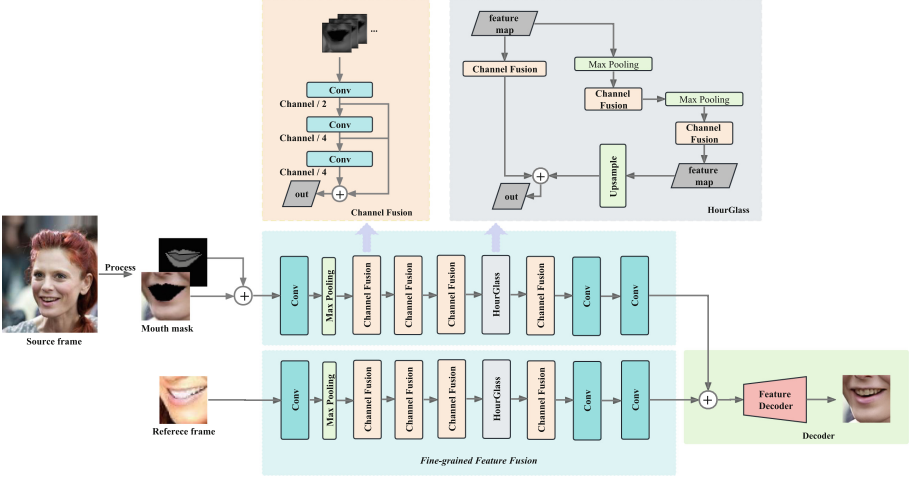
Fig. 2. Pipeline of our HDTR-Net. HDTR-Net consists of two parallel Fine-Grained Feature Fusion (FGFF) modules in the blue rectangles and a Decoder module in the green rectangle. Two important components are Channel Fusion (CF) in the orange rectangular and HourGlass in the grey rectangular included in FGFF. FGFF extracts fine-grained texture features from the processed source image and the reference image, then concatenates the feature map to the Decoder for the restoration of teeth regions. (Color figure online)

3.1 Fine-Grained Feature Fusion

The blue rectangle in Fig. 2 illustrates the structure of Fine-Grained Feature Fusion (FGFF). Given one source image $I_s \in R^{3 \times H \times W}$, after processing the source image for producing the mouth mask image $I_m \in R^{3 \times H \times W}$ and mouth contour image $I_c \in R^{3 \times H \times W}$, I_m and I_c are fed into the Fine-Grained Feature Fusion (FGFF) module for generating the feature map as:

$$\boldsymbol{f_m} = \mathcal{F}((\boldsymbol{I_m} \oplus \boldsymbol{I_c}); \Theta_m) \tag{1}$$

where $\oplus$ is channel concatenation, after processing the source image, I_m and I_c are concatenated into FGFF, which FGFF module $\mathcal{F}(\cdot; \Theta_m)$ with a set of parameters Θ_m transforms I_m and I_c into another feature map $\boldsymbol{f_m}$.

FGFF contains two important components which are the orange rectangular Channel Fusion (CF) and the grey rectangular HourGlass [21] in Fig. 2. CF takes the input of the feature map and passes through three convolutional layers, with the first convolutional layer output a feature map channel that is one-half of the original channel. The last two convolutional layers output a feature map channel that is one-fourth of the original channel. Finally, the output of each convolutional layer is concatenated at the channel dimension, so that the output of CF remains channel invariant but merges with the feature after multi-layer convolutional. Similarly, HourGlass accepts the input of the feature map,

embedding the max pooling on the basis of CF. After two layers of max pooling and CF, the features with larger weights are saved and further merged. Finally, the output feature map after upsampling and the CF output feature map are concatenated at the channel dimension.

Reference image is fed into the branches below FGFF as:

$$\boldsymbol{f_r} = \mathcal{F}(\boldsymbol{I_r}; \Theta_r) \tag{2}$$

where $\boldsymbol{I_r}$ is the reference image, FGFF $\mathcal{F}(\cdot; \Theta_r)$ with a set of parameters Θ_r transforms I_r into another feature map $\boldsymbol{f_r}$.

3.2   Decoder

Decoder focuses on repairing the mouth mask images utilizing two FGFF module results as:

$$\boldsymbol{I_o} = \mathcal{D}((\boldsymbol{f_m} \oplus \boldsymbol{f_r}); \Theta_d) \tag{3}$$

where $\oplus$ is channel concatenation, decoder $\mathcal{D}(\cdot; \Theta_d)$ with a set of parameters Θ_d transforms $\boldsymbol{f_m}$ and $\boldsymbol{f_r}$ into the output $\boldsymbol{I_o}$.

3.3   Loss Function

In the training stage, we use three kinds of loss functions to train HDTR-Net, containing reconstruction loss, perception loss [11], and GAN loss [17].

GAN Loss. Frame discriminator predicts the probability whether the generated frame is comparable to ground truth, resulting in the loss as:

$$\mathcal{L}_{GAN} = \mathcal{L}_D + \mathcal{L}_G \tag{4}$$

where

$$\mathcal{L}_D = \frac{1}{2} E \left(D \left(I_g \right) - 1 \right)^2 + \frac{1}{2} E \left(D \left(I_o \right) - 0 \right)^2 \tag{5}$$

where G represents HDTR-Net and D denotes the discriminator, $\boldsymbol{I_g}$ represents the ground truth image, and $\boldsymbol{I_o}$ represents the teeth restoration image.

Reconstruction Loss. To ensure the coherence of image color and mouth shape, we use L1 loss and L2 loss to reconstruct the mouth region as:

$$\mathcal{L}_{rec}(I_g, I_o) = \|I_g - I_o\|_1 + \|I_g - I_o\|_1^2 \tag{6}$$

Perception Loss. In order to make the generated image have a more natural appearance, we use perception loss to capture the high-level feature differences between the generated image and the ground truth. We calculate the perception loss in $\boldsymbol{I_g}$ and $\boldsymbol{I_o}$ by pre-training the VGG network [3], and the perception loss is formulated as:

$$\mathcal{L}_{\text{perc}} (I_g, I_o) = \|\phi (I_g) - \phi (I_o)\|_2^2 \tag{7}$$

where ϕ is a feature extraction network.

The overall loss is formulated as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{GAN} + \lambda_2 \mathcal{L}_{perc} + \lambda_3 \mathcal{L}_{rec} \tag{8}$$
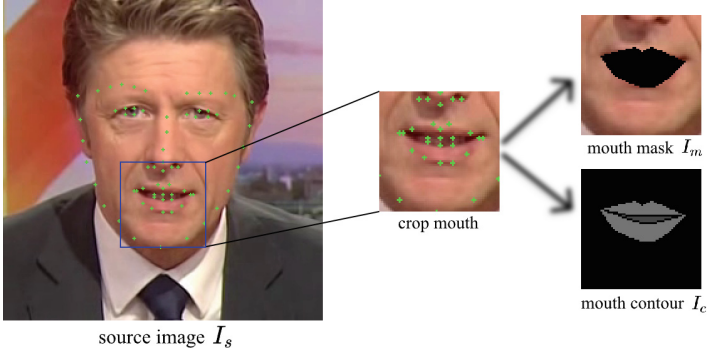
Fig. 3. Details of data processing. Firstly, face landmarks are extracted from source image I_s, then based on the four key points located at the left corner of the mouth, the right corner of the mouth, the tip of the nose, and the jaw, we can determine the boundaries of the mouth region. Using these boundary points, we crop the mouth region from the source image. After processing the cropped mouth region, mouth masks I_m and mouth contour I_c are obtained.

4   Experiment

In this section, we first detail datasets and metrics, comparison methods, and implementation details in our experiment. Then, we show the teeth restoration results of our method. Next, we carry out qualitative and quantitative comparisons with other state-of-the-art works. Finally, we conduct ablation studies.

4.1   Experimental Settings

Dataset and Metrics. We train the proposed method in the train set with high definition processed dataset LRS2, and evaluate with other competitive approaches in the test sets of two prevalent benchmark datasets: LRS2 [1] and LRW [6]. LRS2 contains more than 1000 speakers, with nearly 150,000 instances of words captured and 63,000 different words due to unlimited sentences at the time of capture. LRW selects the 500 most frequent words and clips the speaker's speech, resulting in over 1000 speakers and over 55,000,000 instances of speech.

During the testing phase of our method, there is no ground truth available for direct comparison. To assess the effectiveness of our method, we employ eight image sharpness evaluation metrics: Brenner, Laplacian, SMD, SMD2, Variance, Energy, Vollath, and Entropy. These metrics are chosen to measure the quality of the restored images, aiming to align with human subjective perception. In addition, we measure the time consumption to repair the images.

Comparison Methods. We compare our method with super-resolution based face restoration methods, including ESRGAN [33], GFPGAN [31], and Real-ESRGAN [32]. Real-ESRGAN is an enhanced version of the ESRGAN method.
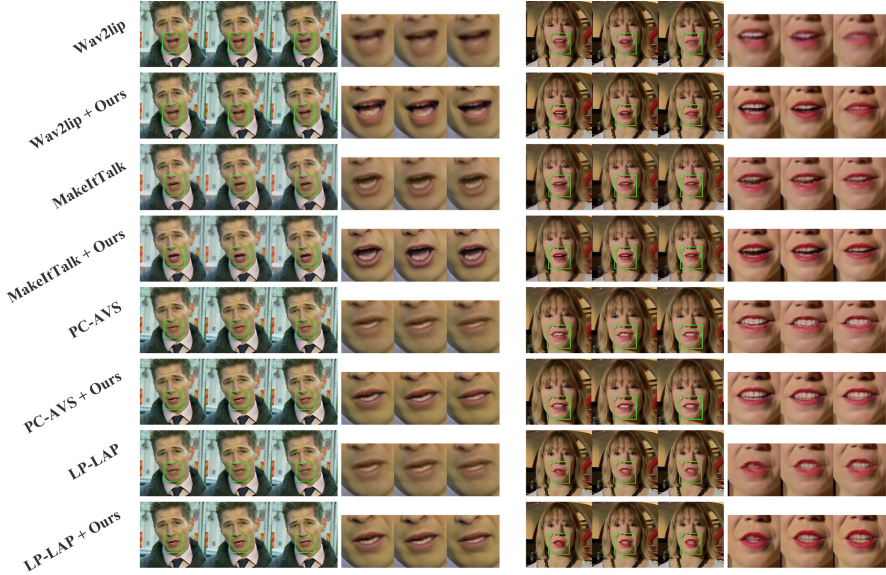
Fig. 4. Our method is adaptable to arbitrary Talking Face Generation methods, and restores the teeth region to obtain high-definition teeth.

These methods input images of different resolutions into the generator to output $1\times$ and multiples of high-resolution enlarged images, but our method's input and output are all 96×96 resolution. To ensure a fair comparison, we maintain the same input and output resolution for all the comparison experiments, the same as our method.

Implementation Details. Data processing is shown in Fig. 3. Given source image $I_s \in R^{3 \times H \times W}$, we begin by extracting face landmarks and cropping the mouth region using four key points: the left corner and the right corner of the mouth, the tip of the nose, and the jaw. Then we obtain a rectangular region containing the mouth and surroundings. Using the extracted face landmarks, we construct a mouth mask $I_m \in R^{3 \times H \times W}$ and mouth contour $I_c \in R^{3 \times H \times W}$ using the key points of the lips. To ensure consistency, $I_m \in R^{3 \times H \times W}$ and $I_c \in R^{3 \times H \times W}$ are aligned to a resolution of 96×96 pixels, focusing on the mouth region. These aligned frames are then input into HDTR-Net.

During training, HDTR-Net takes three inputs: one masked mouth image $I_m \in R^{3 \times 96 \times 96}$, one mouth contour image $I_c \in R^{3 \times 96 \times 96}$, and one randomly selected image $I_f \in R^{3 \times 96 \times 96}$ from the training dataset as a reference frame. We use Adam optimizer [13] with a default setting to optimize HDTR-Net. The learning rate is set to 0.0001. The batch size is set to 12 on one A100 GPU.

Table 1. Quantitative comparisons with the state-of-the-art methods on image quality.

Metrics	Time ↓	Brenner ↑	Laplacian ↑	SMD ↑	SMD2 ↑	Variance ↑	Energy ↑	Vollath ↑	Entropy ↑
ESRGAN	0.0597	2003566	81.38	**163871**	193160	33401059	725269	31201016	4.56
GFPGAN	1.2332	2302120	140.18	72312	240881	6318184	983560	5108233	4.56
Real-ESRGAN	0.0622	2029371	75.15	66755	199590	6190525	772949	5069267	4.55
Ours	**0.0187**	**3121676**	102.28	85607	**331483**	**8990130**	**1264451**	**7667062**	**4.75**

Table 2. Quantitative results of ablation study.

Metrics	Brenner ↑	Laplacian ↑	SMD ↑	SMD2 ↑	Variance ↑	Energy ↑	Vollath ↑	Entropy ↑
w/o CF	2101920	68.18	68294	253783	6392508	923189	6912671	4.21
w/o ref FGFF	2803462	86.12	74029	248926	7183411	871762	6727348	4.26
w/o percep loss	2003462	70.12	78922	272609	7310412	971504	7027348	4.31
Ours	**3121676**	**102.28**	**85607**	**331483**	**8990130**	**1264451**	**7667062**	**4.75**

4.2 Experimental Results

Restoration Results. The results of teeth restoration are shown in Fig. 4. Our method displays the effectiveness of the restored teeth region in four Talking Face Generation methods, including Wav2lip [22], MakeItTalk [43], PC-AVS [42], and IP-LAP [39]. It can be seen that both sharpness and details are obtained after restoring teeth. Wav2lip and MakeItTalk generate blurred areas of the teeth region. After applying our method, it is clear that our method has significant improvements in terms of sharpness, color accuracy, and level of detail in the resulting images. Although PC-AVS and LP-LAP methods are able to generate satisfactory teeth restoration results, our method further improves the clarity, texture, and color of the tooth region. In addition, our method preserves both frame coherence and lip-synchronization, which is friendly handling of every frame in talking face generation videos.

To further verify the teeth restoration robustness for arbitrary talking face generation methods, we evaluate our method in side face talking face videos and show the restoration results in Fig. 5. We observe that our model produces both sharpness and color teeth details.

Quantitative Results. We compare our model with three recent state-of-the-art methods, including ESRGAN [33], GFPGAN [31], and Real-ESRGAN [32]. Table 1 shows the quantitative results of our method and its competitors. Although ESRGAN obtains the highest variance while our method has a second performance, ours surpassed the state-of-the-art ones among other clarity metrics. With the input and output resolutions kept the same, our method achieves an impressive inference time of only 0.018 s per frame. This makes it more than three times faster than the fastest super-resolution restoration method available.

Fig. 5. Visualization results in side-faced speaker. The first and third rows are the original talking face-generated frames, and the second and fourth rows show the results of the dental restoration applied with our method. It can be clearly seen that our model produces both sharpness and color teeth details in a side-faced speaker, demonstrating the effectiveness and robustness of our method.

Qualitative Comparisons. To present the superiority of our method, we provide generated samples compared with ESRGAN, GFPGAN, and Real-ESRGAN. As shown in Fig. 6, face restoration based on super-resolution is not well repaired in frame coherence and tooth texture details, our method has better performance visually in teeth texture details.

4.3   Ablation Study

In order to validate the effect of each component of our method, we conduct ablation study experiments for our HDTR-Net. Specifically, we set 2 conditions: (1) w/o CF: we replace channel fusion with convolution. (2) w/o branches below FGFF module: we remove the branches below FGFF in HDTR-Net. (3) w/o percep loss: we remove perceptual loss in the training stage.

   Table 2 illustrates the qualitative results of ablation experiments. In our condition of Ours w/o reference FGFF module, it is clear to see a significant reduction in results without the reference FGFF module. In our condition of Ours w/o perceptual loss, we also observe a decline in all of our evaluation metrics. We argue that adding CF and ref FGFF can obtain rich structural information and generate images with more high-frequency information. In addition, each component operates effectively in our method.
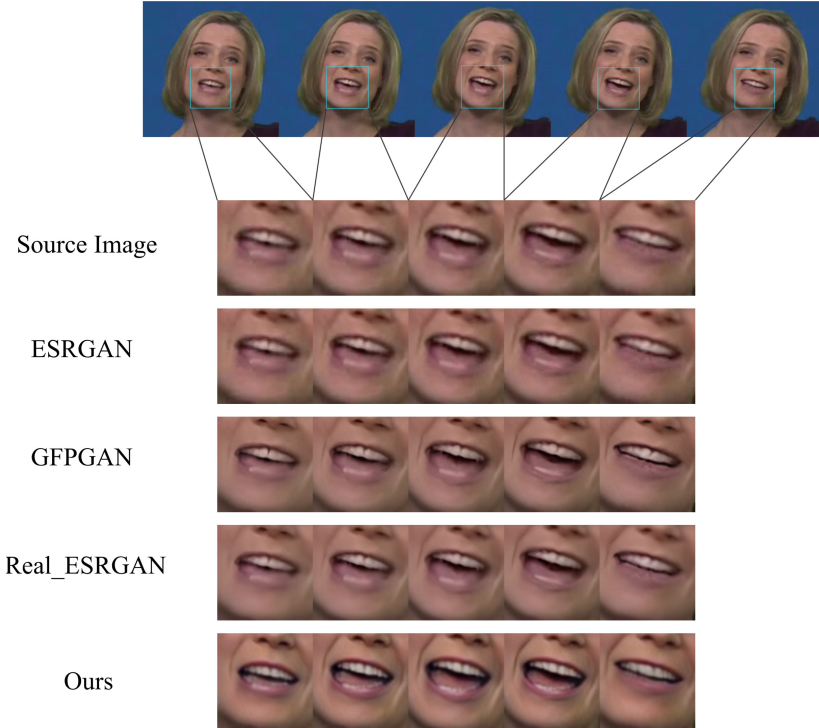
Source Image

ESRGAN

GFPGAN

Real_ESRGAN

Ours

Fig. 6. Compared to face restoration methods based on super-resolution, the visualization results indicate that our method outperforms significantly in teeth texture detail and color parts.

5   Conclusion

In this paper, we propose a real-time High-Definition Teeth Restoration Network (HDTR-Net), including two parallel Fine-Grained Feature Fusion modules and a Decoder module, to realize rich detail and texture in and around the teeth. The Fine-Grained Feature Fusion module is designed to merge low-level edge features and deep-level semantic features in feature dimensions to preserve the image refinement feature. In the inference stage, for two parallel FGFFs, the frame-by-frame guide input of the reference FGFF is capable of repairing teeth while ensuring frame coherence. The Decoder module is designed to merge the extracted feature maps from FGFF and restore the teeth region. With the combination of Fine-Grained Feature Fusion and Decoder, our method preserves more textural details and high-frequency information. Extensive qualitative and quantitative experiments have validated the performance of our method in arbitrary talking face generation methods without suffering lip synchronization and frame coherence. Compared to the super-resolution based face restoration methods, our inference speed is three times faster or even higher.

References

1. Afouras, T., Chung, J.S., Senior, A.W., Vinyals, O., Zisserman, A.: Deep audio-visual speech recognition. IEEE Trans. Pattern Anal. Mach. Intell. **44**(12), 8717–8727 (2022)

2. Cao, Q., Lin, L., Shi, Y., Liang, X., Li, G.: Attention-aware face hallucination via deep reinforcement learning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 1656–1664. IEEE Computer Society (2017)

3. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: Valstar, M.F., French, A.P., Pridmore, T.P. (eds.) British Machine Vision Conference, BMVC 2014, Nottingham, UK, 1–5 September 2014. BMVA Press (2014)

4. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. CoRR abs/1905.03820 (2019)

5. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: FSRNet: end-to-end learning face super-resolution with facial priors. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 2492–2501. Computer Vision Foundation/IEEE Computer Society (2018)

6. Chung, J.S., Senior, A.W., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 3444–3453. IEEE Computer Society (2017)

7. Dogan, B., Gu, S., Timofte, R.: Exemplar guided face image super-resolution without facial landmarks. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, 16–20 June 2019, pp. 1814–1823. Computer Vision Foundation/IEEE (2019)

8. Eskimez, S.E., Maddox, R.K., Xu, C., Duan, Z.: Generating talking face landmarks from speech. In: Deville, Y., Gannot, S., Mason, R., Plumbley, M.D., Ward, D. (eds.) LVA/ICA 2018. LNCS, vol. 10891, pp. 372–381. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93764-9_35

9. Huang, H., He, R., Sun, Z., Tan, T.: Wavelet-SRNet: a wavelet-based CNN for multi-scale face super resolution. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017, pp. 1698–1706. IEEE Computer Society (2017)

10. Jamaludin, A., Chung, J.S., Zisserman, A.: You said that?: synthesising talking faces from audio. Int. J. Comput. Vis. **127**(11–12), 1767–1779 (2019)

11. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43

12. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 1646–1654. IEEE Computer Society (2016)

13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Conference Track Proceedings (2015)

14. Kolouri, S., Rohde, G.K.: Transport-based single frame super resolution of very low resolution face images. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015, pp. 4876–4884. IEEE Computer Society (2015)

15. Kumar, R., Sotelo, J., Kumar, K., de Brébisson, A., Bengio, Y.: ObamaNet: photo-realistic lip-sync from text. arXiv preprint arXiv:1801.01442 (2017)

16. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 105–114. IEEE Computer Society (2017)

17. Lee, C., Cheon, Y., Hwang, W.: Least squares generative adversarial networks-based anomaly detection. IEEE Access **10**, 26920–26930 (2022)

18. Li, X., Li, W., Ren, D., Zhang, H., Wang, M., Zuo, W.: Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020, pp. 2703–2712. Computer Vision Foundation/IEEE (2020)

19. Li, X., Liu, M., Ye, Y., Zuo, W., Lin, L., Yang, R.: Learning warped guidance for blind face restoration. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11217, pp. 278–296. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_17

20. Nagrani, A., Chung, J.S., Albanie, S., Zisserman, A.: Disentangled speech embeddings using cross-modal self-supervision. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, 4–8 May 2020, pp. 6829–6833. IEEE (2020)

21. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29

22. Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.V.: A lip sync expert is all you need for speech to lip generation in the wild. In: Chen, C.W., et al. (eds.) MM 2020: The 28th ACM International Conference on Multimedia, Virtual Event/Seattle, WA, USA, 12–16 October 2020, pp. 484–492. ACM (2020)

23. Sønderby, C.K., Caballero, J., Theis, L., Shi, W., Huszár, F.: Amortised MAP inference for image super-resolution. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017, Conference Track Proceedings. OpenReview.net (2017)

24. Song, Y., Zhu, J., Li, D., Wang, A., Qi, H.: Talking face generation by conditional recurrent adversarial network. In: Kraus, S. (ed.) Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, 10–16 August 2019, pp. 919–925. ijcai.org (2019)

25. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Trans. Graph. **36**(4), 95:1–95:13 (2017)

26. Toshpulatov, M., Lee, W., Lee, S.: Talking human face generation: a survey. Expert Syst. Appl. 119678 (2023)

27. Vougioukas, K., Petridis, S., Pantic, M.: End-to-end speech-driven realistic facial animation with temporal GANs. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, 16–20 June 2019, pp. 37–40. Computer Vision Foundation/IEEE (2019)
28. Vougioukas, K., Petridis, S., Pantic, M.: Realistic speech-driven facial animation with GANs. Int. J. Comput. Vis. **128**(5), 1398–1413 (2020)
29. Wang, G., Zhang, P., Xie, L., Huang, W., Zha, Y.: Attention-based lip audio-visual synthesis for talking face generation in the wild. CoRR abs/2203.03984 (2022)
30. Wang, J., Qian, X., Zhang, M., Tan, R.T., Li, H.: Seeing what you said: talking face generation guided by a lip reading expert. CoRR abs/2303.17480 (2023)
31. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, 19–25 June 2021, pp. 9168–9178. Computer Vision Foundation/IEEE (2021)
32. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-ESRGAN: training real-world blind super-resolution with pure synthetic data. In: IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, 11–17 October 2021, pp. 1905–1914. IEEE (2021)
33. Wang, X., et al.: ESRGAN: enhanced super-resolution generative adversarial networks. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11133, pp. 63–79. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11021-5_5
34. Xu, X., Sun, D., Pan, J., Zhang, Y., Pfister, H., Yang, M.: Learning to super-resolve blurry face and text images. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017, pp. 251–260. IEEE Computer Society (2017)
35. Yu, X., Fernando, B., Hartley, R., Porikli, F.: Super-resolving very low-resolution face images with supplementary attributes. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 908–917. Computer Vision Foundation/IEEE Computer Society (2018)
36. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 318–333. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_20
37. Zhang, Z., Ding, Y.: Adaptive affine transformation: a simple and effective operation for spatial misaligned image generation. In: Magalhães, J., Bimbo, A.D., Satoh, S., Sebe, N., Alameda-Pineda, X., Jin, Q., Oria, V., Toni, L. (eds.) MM 2022: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022, pp. 1167–1176. ACM (2022)
38. Zhang, Z., Hu, Z., Deng, W., Fan, C., Lv, T., Ding, Y.: DINet: deformation inpainting network for realistic face visually dubbing on high resolution video. CoRR abs/2303.03988 (2023)
39. Zhong, W., et al.: Identity-preserving talking face generation with landmark and appearance priors. CoRR abs/2305.08293 (2023)
40. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q.: Learning face hallucination in the wild. In: Bonet, B., Koenig, S. (eds.) Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 25–30 January 2015, Austin, Texas, USA, pp. 3871–3877. AAAI Press (2015)

41. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, 27 January–1 February 2019, pp. 9299–9306. AAAI Press (2019)
42. Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z.: Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, 19–25 June 2021, pp. 4176–4186. Computer Vision Foundation/IEEE (2021)
43. Zhou, Y., Li, D., Han, X., Kalogerakis, E., Shechtman, E., Echevarria, J.: Makeittalk: speaker-aware talking head animation. CoRR abs/2004.12992 (2020)
44. Zhu, H., Huang, H., Li, Y., Zheng, A., He, R.: Arbitrary talking face generation via attentional audio-visual coherence learning. In: Bessiere, C. (ed.) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, pp. 2362–2368. ijcai.org (2020)
45. Zhu, S., Liu, S., Loy, C.C., Tang, X.: Deep cascaded bi-network for face hallucination. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 614–630. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_37