2023 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-6891-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICME55011.2023.00049

CPNet: Exploiting CLIP-based Attention Condenser and Probability Map Guidance for High-fidelity Talking Face Generation

Jingning Xu[1,2], Benlai Tang[2*], Mingjie Wang[3], Minghao Li[2], Meirong Ma[2]
[1]School of Software Engineering, Tongji University, Shanghai, China
[2]Department of AI Technology, Transsion, China
[3]School of Sience, Zhejiang Sci-Tech University, China

***Abstract*—Recently, talking face generation has drawn ever-increasing attention from the research community in computer vision due to its arduous challenges and widespread application scenarios, *e.g.* movie animation and virtual anchor. Although persevering efforts have been undertaken to enhance the fidelity and lip-sync quality of generated talking face videos, there is still large room for further improvements of synthesis quality and efficiency. Actually, these attempts somewhat ignore the explorations of fine-granularity feature extraction/integration and the consistency between probability distributions of landmarks, thereby recurring the issues of local details blurring and degraded fidelity. To mitigate these dilemmas, in this paper, a novel CLIP-based Attention and Probability Map Guided Network (CPNet) is delicately designed for inferring high-fidelity talking face videos. Specifically, considering the demands of fine-grained feature recalibration, a clip-based attention condenser is exploited to transfer knowledge with rich semantic priors from the prevailing CLIP model. Moreover, to guarantee the consistency in probability space and suppress the landmark ambiguity, we creatively propose the density map of facial landmark as auxiliary supervisory signal to guide the landmark distribution learning of generated frame. Extensive experiments on the widely-used benchmark dataset demonstrate the superiority of our CPNet against state of the arts in terms of image and lip-sync quality. In addition, a cohort of studies are also conducted to ablate the impacts of the individual pivotal components.**

***Index Terms*—Talking Face Generation, CLIP, Channel-wise Recalibration, Density Map, Probability Space**

I. Introduction

Inspired by the roaring success of Convolutional Neural Networks(CNNs) during recent decades, the task of talking face generation has caught increasing attention form the research community in the realm of computer vision [1], [2]. Given any speech signal inputs, talking face generation hammers at synthesizing high-quality and realistic face videos, *i.e.* it converts the speech contents to corresponding visual signals. Thanks to fascinating application scenarios, such as movie animation, virtual computer games and virtual anchor, *etc.*, a series of pioneering attempts have been made with the goal of consistently boosting the synthesis performance and efficiency via sophisticated schemes of representation learning [3]–[7]. Albeit the impressive improvements, there is still large room for enhancing image synthesis and lip-sync quality simultaneously [8].

To realize the expectation of obtaining realistic talking face video, several existing algorithms [3], [6], [9] resort to two-stage strategies. They commonly decompose the holistic problem into two sub-stages: *landmark extraction* and *mapping relation learning*. However, the intermediate landmark prediction easily introduces ambiguity into the subsequent generator while suffering form the lack of detailed information (*e.g.* skin texture and background scenarios), thereby resulting in the higher requirements of the recording environment of the data [5]. Collecting such type of data is tedious and arduous in practical applications. Hence, a train of approaches [3], [5], [6], [10] refine the generation procedure and probe into the protocol that feeds 2D landmark-equipped reference image into the pipeline. This refinement allows the networks to further enrich texture and background hints. Despite more detailed cues, these approaches are prone to generate stiff and dull expressions since they rely on a single reference image which involves extremely limited information on facial movement and texture details. To delve into the application of auxiliary signals, Yu *et al.* [9] leverage Canny edge detector to mine geometric hints on hair, clothing and background. However, it is incompetent to endow the model with multi-scale or multi-level property, thereby failing to capture fine-granularity representations.

The inconsistency between landmark inputs and generated faces is another intractable problem. It is worth noting that landmark is incapable of providing lip/teeth details and therefore the inadequate information easily lowers consistency between ground-truth landmarks and predicted mouth regions [8]. To attenuate this issue, studies [1], [8], [11] propose audio aggregation module and attach it to the primary generation stem. However, the heterogeneity caused by multifarious modals hinders the better aggregation of image and audio features, and requires more sophisticated learning architectures or mechanisms. Although a powerful offline lip-sync discriminator is adopted in Wav2Lip [11] as a type of auxiliary signals to assist in the optimization of generator, speaker-specific models provided by Wav2Lip present issues of blurring faces and inconsistent textures [12]. Additionally, approaches [8], [9] try to strengthen the stability of generation from the perspective of inter-frame smoothness, but the inherited inconsistency problem is overlooked.

1945-788X/23/$31.00 ©2023 IEEE
DOI 10.1109/ICME55011.2023.00049

The aforementioned drawbacks stimulate our explorations to improve the image fidelity and lip-sync quality while tackling the inconsistency between landmark ground truth and predicted face frame. In this paper, we propose a Clip-based attention and Probability map guided Network (CPNet) for high-fidelity talking face generation. In specific, inspired by high-efficiency feature reuse in pattern of dense connections [13], we marry our generation backbone with the property of dense feature reuse to mine multi-scale and multi-level semantics. Furthermore, to enrich fine-granularity representations and absorb higher-level priors, we delicately exploit a clip-based [14] attention condenser to transfer knowledge with sufficient multi-modal semantic cues from CLIP and recalibrate the intermediate feature channels of our generator. Last but by no means least, motivated by the prevalence of density map in crowd counting task [15]–[17], we novelly present probability map of landmarks for constraining the consistency between the generated face frame and the landmark groundtruths in probability space instead of inchoate pixel-wise Euclidean distance. The introduction of probability map is also beneficial for imposing landmark distribution learning on our CPNet.

In a nutshell, our main contributions are fourfold.

- **In use:** In this paper, we propose a novel Clip-based attention and Probability map guided Network (CPNet) for effective and high-fidelity talking face generation.
- **Fine granularity:** A *clip-based attention mechanism* is delicately designed to extract fine-granularity representations by transferring priors with rich semantic cues from the prevailing CLIP paradigm.
- **Consistency:** We creatively propose a type of constraints, *probability map*, to guarantee the consistency or smoothness between the generated frames and the probability distributions of the ground-truth landmarks.
- **Performance:** Extensive experiments and ablation studies demonstrate the superiority of our proposed framework in terms of the fidelity and lip-sync quality.

II. Related Work

A. Landmark-to-Face Video Generation

Benefiting from the booming of GAN-based video synthesis, the majority of existing studies attempt to regress the video frames from facial landmark inputs for talking face generation via GAN-based methods [3], [6], [9], [18]. The work [19] carefully designs a rendering framework based on texture searching and selection to synthesize Obama videos with about 17 hours footage. U-Net and implicit condition are adopted in the study [18] to draw an outline of mouth on the cropped input scenes. This strategy makes the rendered mouth region more seamless with the top half part in the target video. These approaches attempt to regress mouth regions associated with speech content. In contrast, several studies employ reference images to assist in the generation under different scenarios [3], [6], [9]. Wherein, Kesim *et al.* [6] and AnyoneNet [3] utilize a reference image from a specified scenario to provide scenario details and speaker texture characterization for fine-grained generation. Yu *et al.* [9] adopt a pix2pixHD to generate the initial frame, and treat foregoing generated images as the input for subsequent generation to guide the learning of robust representations.

B. CLIP-based Knowledge Transfer

Recently, the advent of large-scale visual-language pretraining models reveals their powerful capability of representing semantically rich and high-level visual concepts through natural language supervision. Wherein, the most representative and popular framework is Contrastive Language-Image Pretraining (CLIP) [14] and a vast number of clip-based methods are in vogue in computer vision as it contains abundant multimodal knowledge for a variety of down-stream visual tasks. In the field of image generation, several models [20], [21] make attempts to transfer semantically-rich priors from clip, and attains impressive results. Studies [22] use pre-trained clip to model relationship between image and input text, and illustrate that the latent space of clip is capable of semantically modifying images by moving along the dimension of text embedding. The work [21] endeavors to introduce a CLIP latent residual mapper that is trained for a specific text. By doing so, the hybrid collaboration of strong generation ability of StyleGAN [23] and the extraction of extraordinary visual concepts is achieved. Albeit the great potentials of CLIP, how to transfer priors from CLIP for capturing fine-granularity representations in the realm of image generation is still a burning problem.

III. Methodology

A. Densely-connected Generation Backbone

Following existing state-of-the-art image translation methods [3], [5], [18], the commonly-used pix2pix is adopted as our backbone to regress the realistic pixel-level images. Analogous to the model [5], we feed seven consecutive frames involving the current one frame and three prior frames into our backbone rather than inputting current single scene. Treating a batch of sequential frames as input contributes to model inter-frame relationships. To expand the diversities of receptive fields and semantic levels for intermediate features, the pattern of dense connection [13] is employed in our backbone to reuse features from all preceding transition layers. Apart from the re presentation richness, another merit of dense connection is that it can alleviate gradient vanishing and speed up the convergence of the model. It implies that for any layer x^l in transition layer or decoder, the input is modified as:

$$x^l = x^l + \sum_{i=1}^{3} Pool(H_i^l(e_i)) \tag{1}$$

where e_i is the output of the $i_t h$ layer of encoder, H_i^l is a convolution operation with kernel size 1x1 and $Pool(\cdot)$ is an adaptive pooling operation to perform multi-scale feature map size alignment.

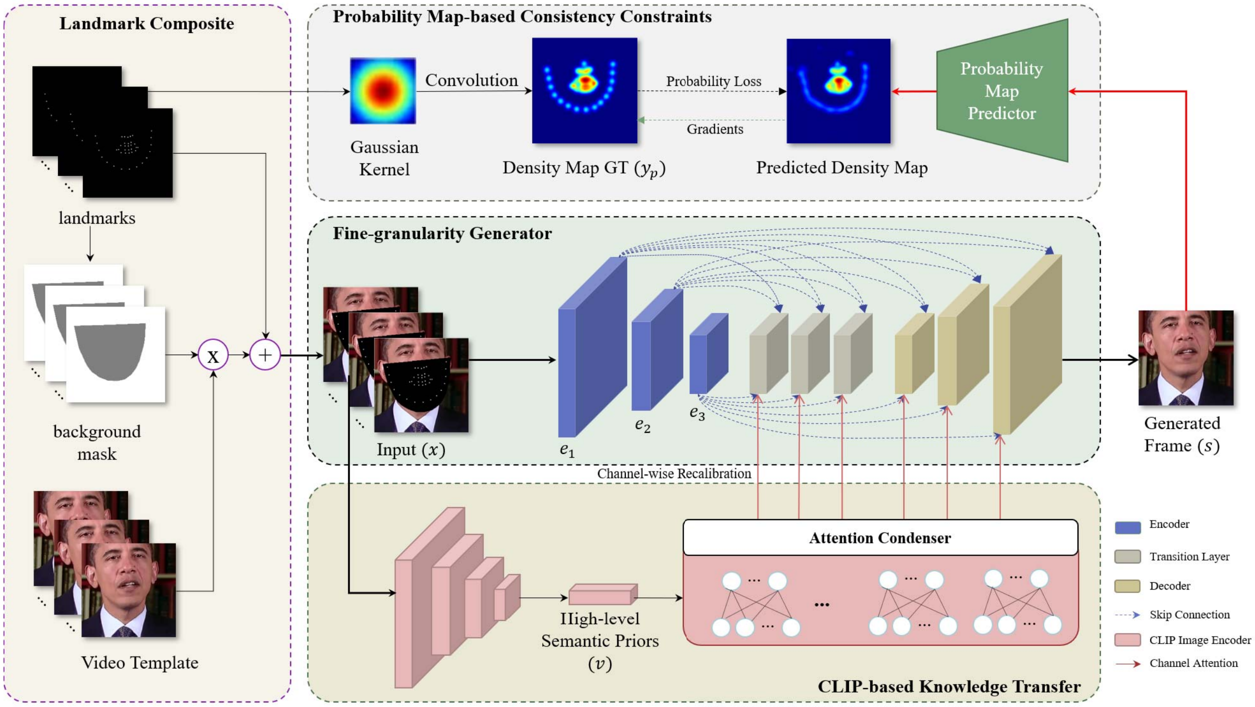

Fig. 1. The overview of the proposed CPNet towards exploiting CLIP-based attention condenser and probability map guidance for high-fidelity talking face generation, which is composed of three modules that includes densely-connected Generator, CLIP-based Knowledge Transfer and Probability Map Predictor.

B. CLIP-based Attention Condenser

As described in Section II, the large-scale pre-trained CLIP is widely used in various image generation tasks thanks to its rich semantic priors through multi-modal contrastive learning. Here we reload the weights of ViT part in CLIP and take the facial landmark image of the current frame as input of clip-based knowledge transfer module. And then ViT stem outputs a one-dimensional vector as the high-level semantic priors. Furthermore, this vector is passed through a SENet [24]-like sub-network to transfer the semantic priors into the channel-wise attention weights. The key difference is that the recalibration weights of our attention condenser are derived from the linear transformation of CLIP latent embedding with sufficient semantic information instead of self-channel features. On top of the adaptively-learned weights, intermediate representations of the generator could be further enhanced by absorbing semantic cues provided by CLIP. And then for each layer x^l in transition layer or decoder, we opt to employ a simple gating mechanism with a sigmoid activation:

$$x^l = F_{scale}(x^l, \sigma(Wv)), \tag{2}$$

where $F_{scale}(\cdot)$ indicates the channel-wise multiplication between the scalar and the feature map, whereas v denotes the CLIP latent codes, W denotes the weights of linear transformation and $\sigma(\cdot)$ represents the sigmoid function.

C. Landmark Probability Map

To guarantee the consistency between the generated talking face and the corresponding ground-truth landmark image in the probability space, we propose a scheme to constraints the predicted frame via auxiliary probability map predictor and density map-based loss function. Probability map is produced by convolving the original facial landmark dot image with a heuristically-defined Gaussian kernel. The results empirically show that when the size of Gaussian kernel is set as 25x25 and sigma is configured as 5 in our task, the probability map-based constraint makes the best positive impacts on the holistic learning procedure of our model.

The probability map predictor is built for the generation of predicted density map of landmarks. The predictor adopts pix2pix network architecture with lightweight parameters and is placed at the end of pipeline. Furthermore, the predictor is not expected to have too much robustness. In other words, the prediction module should be sensitive and discriminative to small changes in the talking face image. Following the design idea of hinge loss, the training objective of this prediction module is set as:

$$l_{dmp} = ||P(I) - y_p||_2 - \lambda||P(I') - P(I)||_2, \tag{3}$$

where P denotes the prediction module and y_p denotes the target probability map and I is the real image, whereas I' represents the image generated by the generator. λ is a adjustment hyperparameter to control the effects of the prediction module.

D. Objective Function

The ultimate training objective is to supervise the optimization of our proposed framework for regressing high-fidelity talking face video with high lip-sync quality. To simplify the subsequent explanations, the notation is defined here. We denote the sequence of consecutive frames input to the generator as x, whereas the corresponding ground truth frame is y and s indicates the generated frame.

Following LSGAN loss [25], we first implement the adversarial loss L_{adv} as:

$$L_{adv} = \mathbb{E}_{x,y}[(D(x,y)-1)^2] + \mathbb{E}_{x,s}[D(x,s)^2], \tag{4}$$

where $\mathbb{E}[\cdot]$ denotes expectation values and D means the discriminator network.

Meanwhile, following the configuration in [26], the perceptual reconstruction loss L_r provides the supervisory signal to impel the generated frames to be close to the ground truths. By introducing perceptual loss term, the quality of predicted images can be improved based on the differences between high-level image representations extracted from pre-trained convolutional neural networks, such as VGGNet [27]. The perceptual perceptual loss l_r is defined as:

$$L_r = \frac{1}{l} \sum_{l} ||\phi^l(s) - \phi^l(y)||_1, \tag{5}$$

where ϕ^l is the l^{th} layer of the pre-trained VGGNet.

Moreover, the video loss L_t is implemented by sequence discriminator D_t to facilitate the video quality, especially for spatio-temporal coherence. Similar to L_{adv}, L_t is defined as:

$$L_t = \mathbb{E}_{x^t,y^t}[(D_t(x^t,y^t)-1)^2] + \mathbb{E}_{x^t,s^t}[D_t(x^t,s^t)^2], \tag{6}$$

where D_t denotes the discriminator network, x^t, y^t, s^t are the corresponding input, ground truth and the generated image, respectively, for consecutive several frames.

To enhance the consistency in probability space, a pivotal probability map loss term is designed here to shorten the distance between the generated image and ground-truth landmark distribution. We denote the probability map predictor as P and the probability map loss l_p is written as:

$$L_p = ||P(s) - P(y)||_2 \tag{7}$$

In summary, the entire loss function of our CPNet is formulated as:

$$L = \lambda_{adv} L_{adv} + \lambda_r L_r + \lambda_t L_t + \lambda_p L_p, \tag{8}$$

where $\lambda_{adv}, \lambda_r, \lambda_t, \lambda_p$ are hyperparameters which are utilized to adjust the influences of multifarious loss terms.

IV. Experiments

A. Datasets and Implementation Details

Datasets. Our CPNet is verified on the prevailing ObamaSet benchmark [19] which is widely chosen to evaluate lip-sync methods. This dataset includes the total of 17 hours videos at 29.97 fps with a resolution of 1280×720. This dataset contains hundreds of scenarios from Obama's weekly speeches with a wide range of clothing types, lighting conditions and backgrounds. We randomly select 200 videos in total of 3 hours from the original dataset. Wherein, 90% of the samples are used as the training set while the rest of 10% is treated as the testing set to testify the performance of talking face generation during the inference phase.

Evaluation Metrics. In our experiments, to evaluate the quality of the synthesized frames, common reconstruction metrics like the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) metrics [28] are used to reflect the visual quality. And the Fréchet Video Distance (FVD) is used to evaluate the spatio-temporal consistency of the generated videos [29]. Besides, "LSE-D" and "LSE-C", proposed in [11], is adopted to evaluate the lip-sync quality of generated videos. The former calculates the distance between the lip and audio representations and the latter metric the average confidence score between the speech and lip movements.

Implementation Details. The weights for the loss terms are empirically fixed as: $\lambda_{adv} = 1, \lambda_r = 5, \lambda_t = 1, \lambda_p = 0.1$. We adopt Adam optimizer with the learning rate of 0.0001 and the momentum of 0.5 to optimize our CPNet for around 100k iterations. During the training phase, frames with the size of 224×224 are centrally cropped around the center of facial region.

B. Comparison with State of the Arts

We reimplement the "landmark to photo-realistic image" module in AnyoneNet [3] and "Edge-to-Video" module in ADEVP [30] as baselines, whose objective is essentially the same as ours. The quantitative results are demonstrated in Table I while the visualization examples are also provided in Figure 2. Both quantitative results and visualization examples demonstrate the superiority of our proposed framework. It can be concluded that our proposed approach is better than existing methods in all aspects of image quality, video consistency, and lip-sync quality.

TABLE I
Comparisons with baseline models on ObamaSet. The proposed method consistently surpasses the existing approaches on all metrics. "↑" means the greater the metric, the better the generation, and "↓" is the opposite.

Method	SSIM ↑	PSNR ↑	FVD ↓	LSE-C ↑	LSE-D ↓
AnyoneNet [3]	0.959	34.0	0.146	0.236	14.98
ADEVP [30]	0.963	34.6	0.113	0.299	14.63
Our approach	**0.973**	**35.6**	**0.060**	**0.401**	**13.66**

C. Ablation Study

1) The Impacts of Individual Components: To better understand each component in our model and verify the effects of them, we carry out various ablation studies on ObamaSet benchmark (see Table II). The results demonstrates that densely-connected generator (*I*) has an impressive contribution to visual quality. Thanks to the introduction of rich semantic information, the image quality and fidelity of the generated results have been greatly improved by exploiting the attention condenser through CLIP latent embedding (*II*). And the probability map constraints of landmark (*III*) also have positive impacts on the lip-sync metrics.

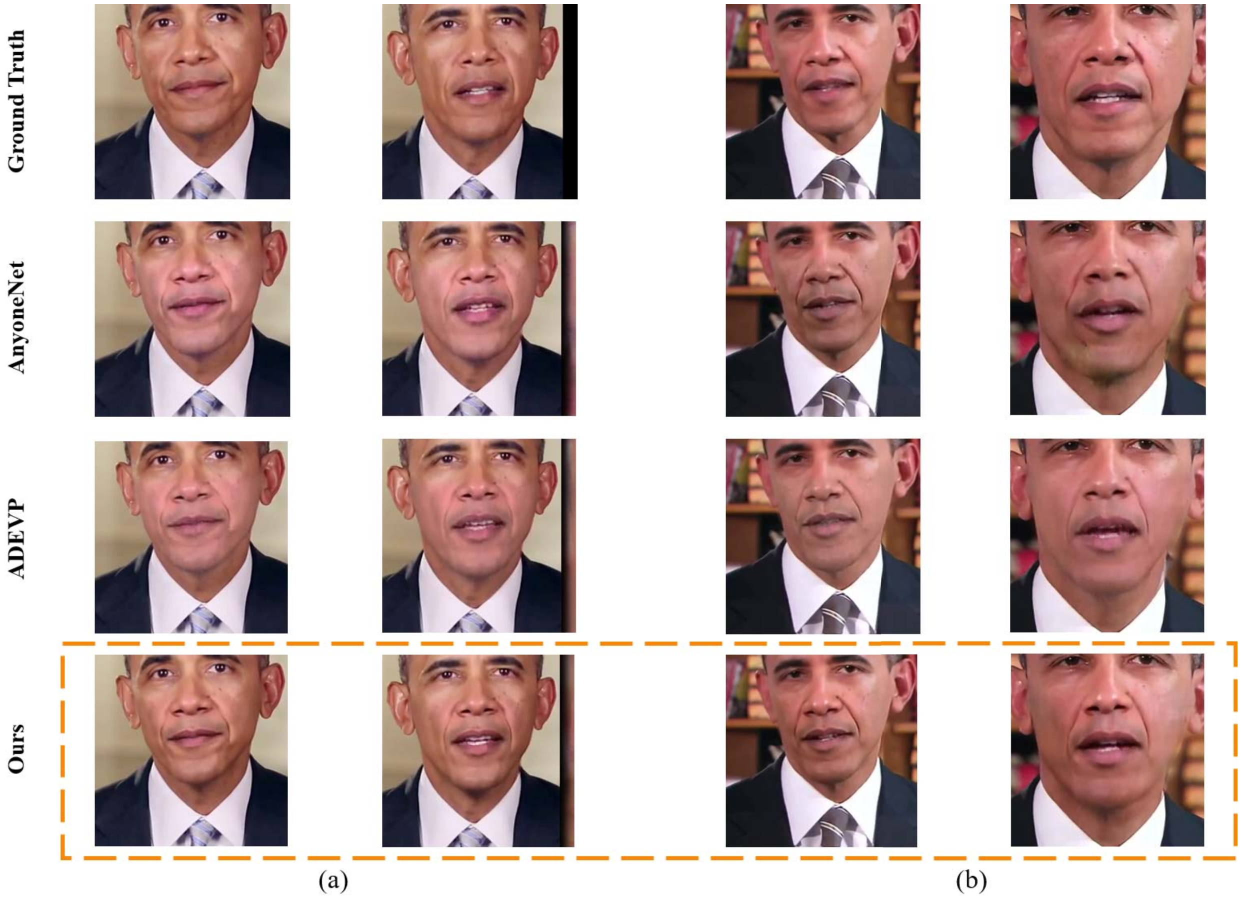

Fig. 2. The visualization comparisons among baselines and our CPNet on two scenarios ((a) and (b)) from ObamaSet dataset. It can be observed that our method produces more realistic results than AnyoneNet [3] and ADEVP [30], especially for the generations of skin and lip textures.

TABLE II

ABLATION STUDIES FOR PROPOSED INDIVIDUAL COMPONENTS: DENSELY-CONNECTED GENERATOR (*I*), CLIP-BASED ATTENTION CONDENSER (*II*) AND PROBABILITY MAP CONSTRAINTS (*III*).

I	*II*	*III*	SSIM ↑	PSNR ↑	FVD ↓	LSE-C ↑	LSE-D ↓
			0.950	32.9	0.157	0.197	15.33
✓			0.961	34.1	0.108	0.23	14.85
✓	✓		0.971	35.4	0.065	0.312	14.21
✓	✓	✓	**0.973**	**35.6**	**0.060**	**0.401**	**13.66**

2) How do Loss Terms Influence the Performance: To further investigate the effects of all loss terms in our objective function on the performance, we conduct extensive experiments to ablate all loss terms in Table III. It can be observed that using only standard GAN loss of l_{adv} to guide the generation is rather terrible. Additionally, l_r is designed to match the features extracted by the VGG network and allows for alleviating the issue of image blurring and bringing remarkable improvements in image quality metrics. Specifically, compared to the results without l_t, the FVD scores are significantly increased and the quality of the generated image is substantially improved with l_t. Finally, the introduction of l_p significantly improves the quality of the generated lip-sync results.

TABLE III

THE IMPACTS OF MULTIFARIOUS LOSS TERMS ON MODEL PERFORMANCE.

Loss Functions	SSIM ↑	PSNR ↑	FVD ↓	LSE-C ↑	LSE-D ↓
L_{adv}	0.966	34.7	0.151	0.252	14.69
$L_{adv} + L_r$	0.970	35.5	0.129	0.288	14.51
$L_{adv} + L_r + L_t$	0.971	35.4	0.065	0.312	14.21
$L_{adv} + L_r + L_t + L_p$	**0.973**	**35.6**	**0.060**	**0.401**	**13.66**

As for the hyperparameters in our loss function, multiple sets of experiments are carried out and we expirically fix the final weights. Taking L_p as an example, we set weight as 0.05, 0.1, 0.5 and 1.0 for four sets of experiments and obtained the relevant results, see Table IV. As shown in Table IV, the best performance is obtained when the value of L_p is 0.1. We hence fix the L_p as value of 0.1. The controlling weights for remaining loss terms are also determined in this way.

TABLE IV

THE EFFECTS OF HYPERPARAMETERS IN OBJECTIVE FUNCTION.

Hyperparameter λ_p	SSIM ↑	PSNR ↑	FVD ↓	LSE-C ↑	LSE-D ↓
1.0	0.965	34.6	0.063	**0.419**	**13.58**
0.5	0.967	34.8	0.066	0.415	13.61
0.1	**0.973**	**35.6**	**0.060**	0.401	13.66
0.05	0.972	35.4	0.061	0.342	14.07

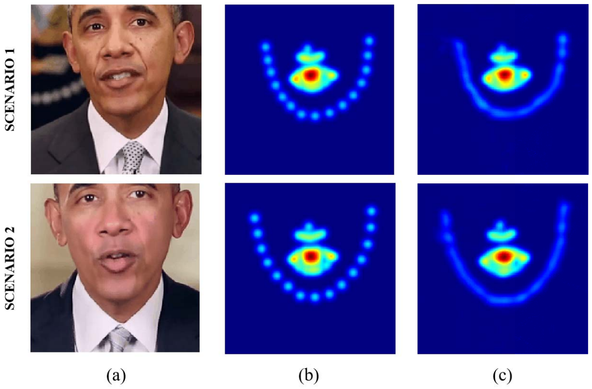

Fig. 3. Probability map prediction results on different scenarios. (a) the generated frames; (b) the ground truths of probability map; (c) the predicted probability maps by our model.

V. Conclusion

In this paper, we exploit clip-based attention condenser and probability map guidance for high-fidelity and lip-sync talking face generation. Specifically, a densely-connected generation backbone is proposed to mine fine-granularity representations with diverse scales and levels. Then, an attention condenser is delicately designed to transfer priors from pre-trained CLIP models to enrich the multi-modal semantic cues. Finally, we devise a new probability map constraint of landmark to guarantee the consistency between generated images and groundtruths in probability space of landmark. Extensive experiments and ablation studies on ObamaSet dataset illustrate the effectiveness and superiority of our proposed CPNet.

Acknowledgment

We would like to express our sincere gratitude to our colleagues of engineering services team who provided essential computational resources during experiments. Additionally, we are also grateful for the support provided by the Science Foundation of Zhejiang Sci-Tech University (ZSTU) under Grant No. 22062338-Y.

References

[1] Wentao Wang, Yan Wang, Jianqing Sun, Qingsong Liu, Jiaen Liang, and Teng Li, "Speech driven talking head generation via attentional landmarks based representation.," in *INTERSPEECH*, 2020.

[2] Aihua Zheng, Feixia Zhu, Hao Zhu, Mandi Luo, and Ran He, "Talking face generation via learning semantic and temporal synchronous landmarks," in *ICPR*, 2021.

[3] Xinsheng Wang, Qicong Xie, Jihua Zhu, Lei Xie, and Odette Scharenborg, "Anyonenet: Synchronized speech and talking head generation for arbitrary persons," *IEEE Transactions on Multimedia*, 2022.

[4] Hyoung-Kyu Song, Sang Hoon Woo, Junhyeok Lee, Seungmin Yang, Hyunjae Cho, Youseong Lee, Dongho Choi, and Kang-wook Kim, "Talking face generation with multilingual tts," in *CVPR*, 2022.

[5] Tianyi Xie, Liucheng Liao, Cheng Bi, Benlai Tang, Xiang Yin, Jianfei Yang, Mingjie Wang, Jiali Yao, Yang Zhang, and Zejun Ma, "Towards realistic visual dubbing with heterogeneous sources," in *ACM MM*, 2021.

[6] Ege Kesim and Engin Erzin, "Investigating contributions of speech and facial landmarks for talking head generation.," in *Interspeech*, 2021.

[7] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *CVPR*, 2021.

[8] Jun Ling, Xu Tan, Liyang Chen, Runnan Li, Yuchao Zhang, Sheng Zhao, and Li Song, "Stableface: Analyzing and improving motion stability for talking face generation," *arXiv preprint arXiv:2208.13717*, 2022.

[9] Lingyun Yu, Jun Yu, Mengyan Li, and Qiang Ling, "Multimodal inputs driven talking face generation with spatial–temporal dependency," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 203–216, 2020.

[10] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *ICCV*, 2019.

[11] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *ACM MM*, 2020.

[12] Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler, "Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization," in *CVPR*, 2021.

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[15] Jia Wan, Qingzhong Wang, and Antoni B Chan, "Kernel-based density map generation for dense object counting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[16] Mingjie Wang, Hao Cai, Xianfeng Han, Jun Zhou, and Minglun Gong, "Stnet: Scale tree network with multi-level auxiliator for crowd counting," *IEEE Transactions on Multimedia*, 2022.

[17] Mingjie Wang, Hao Cai, Yong Dai, and Minglun Gong, "Dynamic mixture of counter network for location-agnostic crowd counting," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 167–177.

[18] Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brébisson, and Yoshua Bengio, "Obamanet: Photo-realistic lip-sync from text," *arXiv preprint arXiv:1801.01442*, 2017.

[19] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–13, 2017.

[20] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," in *CVPR*, 2021.

[21] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or, "Stylegan-nada: Clip-guided domain adaptation of image generators," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–13, 2022.

[22] Nasir Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa, "Clip-mesh: Generating textured meshes from text using pretrained image-text models," *ACM Transactions on Graphics (TOG).*, vol. 3, 2022.

[23] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019.

[24] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[25] X. Mao, Q. Li, H. Xie, Ryk Lau, and S. P. Smolley, "Least squares generative adversarial networks," in *ICCV*, 2017.

[26] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[27] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[28] Alain Hore and Djemel Ziou, "Image quality metrics: Psnr vs. ssim," in *ICPR*, 2010.

[29] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv preprint arXiv:1812.01717*, 2018.

[30] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu, "Audio-driven emotional video portraits," in *CVPR*, 2021.