Implicit Identity Representation Conditioned Memory Compensation Network for Talking Head video Generation

- Supplementary Material -

Fa-Ting Hong
CSE, HKUST
fhongac@cse.ust.hk

Dan Xu*
CSE, HKUST
danxu@cse.ust.hk

1. More implementation details

The keypoint and dense motion estimation follow FOMM [10]. Specifically, we extract each frame from the driving video as a driving image and input it into the MCNet model with the source image. The source image and driving video share the same identity in the training stage, so the sampled driving frame can be used as the ground-truth of a generated source-identity image. To optimize the training objectives, we set $\lambda_{rec} = 10$, $\lambda_{eq} = 10$, $\lambda_{dist} = 10$, and $\lambda_{con} = 10$. The number of keypoints is set to 15, which is the same as that of DaGAN [5]. In the training stage, we employ 8 RTX 3090 GPUs to run the model for 100 epochs in an end-to-end manner, and it costs about 12 hours in total. The number of layers (*i.e.* N) of the encoder and the decoder is both set as 4, and the number of keypoints (*i.e.* K) is set as 15 following [5]. We set the size of the proposed meta memory as $512 \times 32 \times 32$, *i.e.* $C_m = 512$, $H_m = 32$ and $W_m = 32$. In the motion-based warping process, for any $\mathbf{F}_e^i$ that has a different spatial size to the motion flow, we employ bilinear interpolation to adjust the spatial size of the motion flow.

2. Network architecture details of MCNet

The keypoint detector receives an image as input and outputs the K keypoints $\{x_i, y_i\}_{i=1}^K$. The structure of the keypoint detector is illustrated in Fig. 1. Here, we adopt the Taylor approximation as FOMM [10] and DaGAN [5] to compute the motion flow. Thus, the motion estimation is not our focus and we mainly focus on designing our meta memory and its usage for our talking head generation framework.

3. More details on optimization losses

Perceptual Loss $\mathcal{L}_p$. Perceptual loss is a popular objective function in image generation tasks. As introduced in DaGAN [5], a generated image and its ground-truth, *i.e.* the driving image in the training stage, is downsampled to 4 different resolutions (*i.e.* 256×256, 128×128, 64×64, 32×32), respectively. Then we utilize a pre-trained VGG network [12] to extract the features from images at each resolution. To simplify, we denote R_1, R_2, R_3, R_4 as the features of generated images in different resolutions, respectively, and G_1, G_2, G_3, G_4 for the 4 different resolutions of the ground-truth. Then, we measure the $\mathcal{L}_1$ distance between the ground-truth and the generated image by the Perceptual loss defined as follows:

$$\mathcal{L}_p = \sum_{i=1}^{4} \mathcal{L}_1(G_i, R_i) \tag{1}$$

Equivariance Loss $\mathcal{L}_{eq}$. We employ this loss to maintain the consistency of the estimated keypoints in the images after different augmentations. Per FOMM [10], given an image I and its detected keypoints $\{X_i\}_{i=1}^K$ ($X_i \in \mathbb{R}^{1\times 2}$), we first perform a known spatial transformation T on images I and keypoints $\{X_i\}_{i=1}^K$, resulting in a transformed image I_T and transformed keypoints $\{X_i^T\}_{i=1}^K$. Then, we detect keypoints on the transformed image I_T, which are denoted as $(\{X_{I_T,i}\}_{i=1}^K)$. We employ the equivariance Loss on the

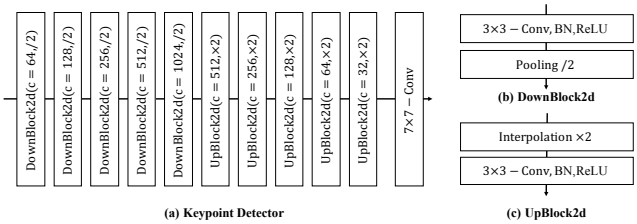

Figure 1: Detailed structure of the keypoint detector. c in each layer indicates the number of output channels.

*Corresponding author

source image and driving image:

$$\mathcal{L}_{eq} = \sum_{i=1}^{K} ||X_i^T - X_{I_T,i}||_1 \quad (2)$$

Keypoint distance loss $\mathcal{L}_{dist}$. We employ the keypoint distance loss as in [5] to penalize the model if the distance between any two keypoints is smaller than a user-defined threshold. Thus, the keypoint distance loss can make the keypoints much less crowded around a small neighbourhood. In one image, for every two keypoints X_i and X_j, we then have:

$$\mathcal{L}_{dist} = \sum_{i=1}^{K}\sum_{j=1}^{K}(1 - \mathbf{sign}(||X_i - X_j||_1 - \alpha)), i \neq j, \quad (3)$$

where the $\mathbf{sign}(\cdot)$ represents a sign function and the α is the threshold of the distance, which is 0.2 in our work.

4. More details on experiments

4.1. Evaluation Metrics

We mainly consider four important metrics that are widely used in the talking head generation field, *i.e.*, AED, ADK, PRMSE, and AUCON. Specifically, **Average euclidean distance (AED)** is an important metric that measures identity preservation in reconstructed video/image. In this work, we use OpenFace [1] to extract identity embeddings from the reconstructed face and the ground truth frame. The MSE loss is used to measure their difference.

Average keypoint distance (ADK). ADK evaluates the difference between landmarks of the reconstructed faces and the ground truth frames. We extract facial landmarks using a face alignment method [2]. We compute an average distance between the corresponding keypoints. Thus, the AKD mainly measures the ability of the pose imitation.

The root mean square error of the head pose angles (PRMSE). In this work, we utilize the Py-Feat toolkit[1] to detect the Euler angles of the head pose, and then evaluate the pose difference between different identities.

The ratio of identical facial action unit values (AUCON). We first utilize the Py-Feat toolkit to detect the action units of the generated face and the driving face. Then we can calculate the ratio of identical facial action unit values as the AUCON metric.

4.2. Additional experimental results

Positional Encoding for keypoints. The positional encoding method shows its strong power in transformers [13, 6, 4] and NeRFs [8, 7, 9]. Therefore, we consider applying the positional encoding function[2] on the keypoints, to produce the implicit identity representation conditioned memory. We show the results in Table.1. From Table 1, we observe that when we apply the position encoding function on keypoints, it cannot bring improvements, and even degrades the model performance if we set the L as 20. Since the keypoints are utilized to estimate the motion flow in the dense motion network, the Euclidean distance between any two keypoints is physically meaningful. Therefore, we suppose that employing the positional encoding on keypoints may affect the motion flow estimation, resulting in an unsatisfactory generation.

Model	SSIM (%) ↑	PSNR ↑	LPIPS ↓	$\mathcal{L}_1$ ↓	AKD ↓	AED ↓
Ous w/ pe(10)	82.4	31.91	0.175	0.0334	1.221	0.107
Ous w/ pe(20)	69.4	30.03	0.269	0.0593	5.544	0.268
MCNet	**82.5**	**31.94**	**0.174**	**0.0331**	**1.203**	**0.106**

Table 1: The results of applying positional encoding function on keypoints. "pe(10)" means that we set the output dimension control factor **L** of positional encoding function as 10, and 20 for "pe(20)".

Model	SSIM (%) ↑	PSNR ↑	LPIPS ↓	$\mathcal{L}_1$ ↓	AKD ↓	AED ↓
MCNet (IICM w/o $\mathbf{F}^i_{proj}$)	82.3	31.89	0.175	0.0336	1.237	0.110
MCNet (IICM w/o keypoints)	82.4	31.93	0.175	0.0333	1.227	0.109
MCNet (MCM w/o f^1_{dc}, f^2_{dc})	82.2	31.89	0.176	0.0336	1.246	0.112
MCNet (single layer)	82.3	31.90	0.175	0.0334	1.235	0.108
MCNet	**82.5**	**31.94**	**0.174**	**0.0331**	**1.203**	**0.106**

Table 2: Ablation studies. 'IICM w/o $\mathbf{F}^i_{proj}$" and "IICM w/o keypoints" represent that IICM does not use the projected feature $\mathbf{F}^i_{proj}$ or keypoints as input (see Fig. **??**), respectively, to encode implicit identity representation. "IICM w/o f^1_{dc}, f^2_{dc}" indicates that we replace the f^1_{dc} and f^2_{dc} with two normal convolution layers to produce the key and the value in MCM.

The input elements in IICM. We also conduct experiments to investigate the usage of intermediate feature $\mathbf{F}^i_{proj}$ ("IICM w/o $\mathbf{F}^i_{proj}$") and keypoints (IICM w/o keypoints). The results are shown in Table 2. The results in the table indicate that these two items are both critical for the generation of the implicit-identity representation conditioned memory bank. We can obtain the best results when we combine them together.

Single layer vs. multiple layers. In our work, we deploy the IICM and MCM in each layer to obtain the best results. Also, we investigate the performance of using IICM and MCM in the first layer only. The results "MCNet (single layer)" show that the single layer can also obtain similar good results, which can verify the effectiveness of our designed memory mechanism.

The dynamic convolution in MCM. Besides, we also con-

[1] https://py-feat.org

[2] Here, we use the implementation of https://github.com/yenchenlin/nerf-pytorch

duct an ablation study on the dynamic convolution layer in the memory compensation module. We can observe that the dynamic convolution layer can contribute to the final performance, especially for the AKD and AED.

Model	SSIM (%) ↑	PSNR ↑	LPIPS ↓	$\mathcal{L}_1$ ↓	AKD ↓	AED ↓
FOMM [10]	77.19	30.71	0.257	0.0513	1.762	0.212
MRAA [11]	78.07	30.89	0.262	0.0511	1.796	0.213
DaGAN [5]	79.02	30.81	0.250	0.0483	1.865	0.341
TPSM [15]	78.22	30.63	0.254	0.0527	1.703	0.210
Ours w/o IICM	78.63	31.02	0.250	0.0481	1.726	0.199
Ours	**79.86**	**31.18**	**0.244**	**0.0470**	**1.699**	**0.186**

Table 3: State-of-the-art comparison on VoxCeleb2 dataset.

Model	SSIM (%) ↑	PSNR ↑	LPIPS ↓	$\mathcal{L}_1$ ↓	AKD ↓	AED ↓
FOMM [10]	76.94	31.87	0.155	0.0363	1.116	0.092
MRAA [11]	79.36	32.32	0.156	0.0331	1.039	0.100
DaGAN [5]	82.29	32.29	0.136	0.0304	1.020	0.083
TPSM [15]	86.05	32.85	0.114	0.0264	1.015	0.072
Ours w/o IICM	85.90	33.03	0.114	0.0243	1.023	0.068
Ours	**86.45**	**33.60**	**0.112**	**0.0238**	**0.998**	**0.064**

Table 4: State-of-the-art comparison on HDTF dataset.

More datasets for evaluation. To fully verify the superiority of our method, we also compare it with other state-of-the-art methods on two other large-scale datasets, *i.e*. VoxCeleb2 [3] and HDTF [14]. We report the results in Table 3 and Table 4. From these two tables, we can observe that our method can still obtain the best results compared with the SOTA methods[3]. These results clearly confirm the superiority of our designed method.

Idenetity Preservation. In this section, we reorganize the voxceleb1 dataset and divide it into a training set and a test set. These two sets have the same identity space. That is, the identities of test videos also appear in the training videos. We select 500 videos as the test set and the rest as the training set. The experimental results are shown in Table 5. We can observe that our method obtains higher performance under the setting of the testing identities as a part of the training corpus. One possible reason is that our global face meta-memory is learned from the identities in the training set. In this way, it can better compensate for the facial details of those seen identities.

Model	SSIM (%) ↑	PSNR ↑	LPIPS ↓	$\mathcal{L}_1$ ↓	AKD ↓	AED ↓
Identities not in the training set	82.5	31.94	0.174	0.0331	1.203	0.106
Identities in the training set	**83.6**	**32.38**	**0.163**	**0.0319**	**1.164**	**0.102**

Table 5: Comparison of different variants on HDTF dataset.

Video generation demo. We also provide several video generation demos to show a more detailed comparison qualitatively with the most competitive methods in the literature. From the demo videos, we can observe that our proposed memory compensation network can compensate the regions that do not appear in the source image, with significantly better results than other methods (*e.g*. the ear region in demo2). These demos are attached in Supplementary Material.

Comparison on tasks in other domains. To better verify the generalization ability of our method, we also train our method on TED-talks dataset [11], because the human body is also symmetrical and highly structured. We report the results in Table 6. From the Table 6, our method still obtain the best results among all the compared methods. This generalization experiment verifies that our meta-memory can learn the symmetrical and structured face information to inpaint the generated image. As shown in Fig. 2, our method yields high-quality body details and learns a memory bank with rich and representative body information, which is very useful for the full-body generation.

Model	$\mathcal{L}_1$ ↓	(AKD ↓, MKR ↓)	AED ↓
FOMM [10]	0.033	(7.07, 0.014)	0.163
MRAA [11]	0.026	(4.01, 0.012)	0.116
TPSM [15]	0.027	(3.39, 0.007)	0.124
Ours	**0.023**	(**2.52**, **0.006**)	**0.101**

Table 6: State-of-the-art comparison on TED-talks dataset.

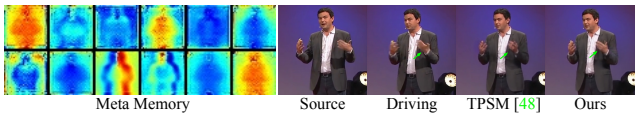

Figure 2: Comparison with TPSM on full-body TED-talks dataset.

Model	SSIM (%) ↑	PSNR ↑	LPIPS ↓	$\mathcal{L}_1$ ↓	AKD ↓	AED ↓
size = 16	82.3	31.83	0.176	0.0338	1.230	0.112
size = 64	82.1	31.75	0.176	0.0343	1.244	0.111
Ours (size = 32)	**82.5**	**31.94**	**0.174**	**0.0331**	**1.203**	**0.106**

Table 7: Ablation study on different memory bank sizes.

The different sizes of the memory bank. We show the ablation studies in Tab. 7 to inverstigate the effectiveness of different sizes of the memory bank. As can be seen in Tab. 7, the size is not sensitive to generation performance. We obtain the best generation results with the size of 32, which is used for all the experiments in the paper.

Meta memory visualization. In this section, we show all the channels of our learned meta-memory in Fig. 3 for better understanding. To better show the details, we also visualize some channels in Fig. 4 in high resolution. These

[3]These compared methods have officially released code for us to test on these two datasets.

visualizations demonstrate the meaningful facial priors are effectively learned in the meta-memory.

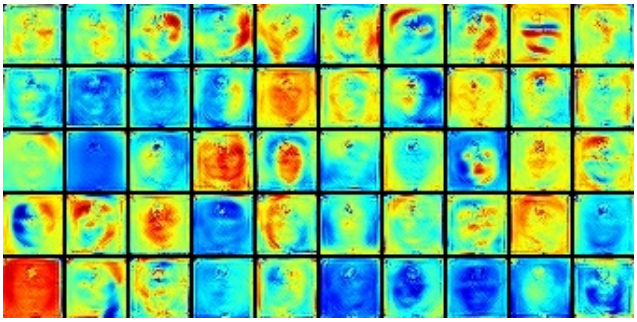

Figure 3: Visualization of randomly selected channels of the meta memory $\mathbf{M}_o$.

More qualitative ablation studies. To better show that our designed implicit identity representation conditioned memory compensation network brings improvements, we present more qualitative results for ablation studies in Fig. 5 and Fig. 6. The effectiveness can be easily observed from the qualitative examples shown in the tables.

References

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *WACV*, 2016. 2

[2] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017. 2

[3] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 3

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[5] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *CVPR*, 2022. 1, 2, 3

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2

[7] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021. 2

[8] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[9] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, 2021. 2

[10] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *NeurIPS*, 2019. 1, 3

[11] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021. 3

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 2

[14] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *CVPR*, 2021. 3

[15] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *CVPR*, 2022. 3

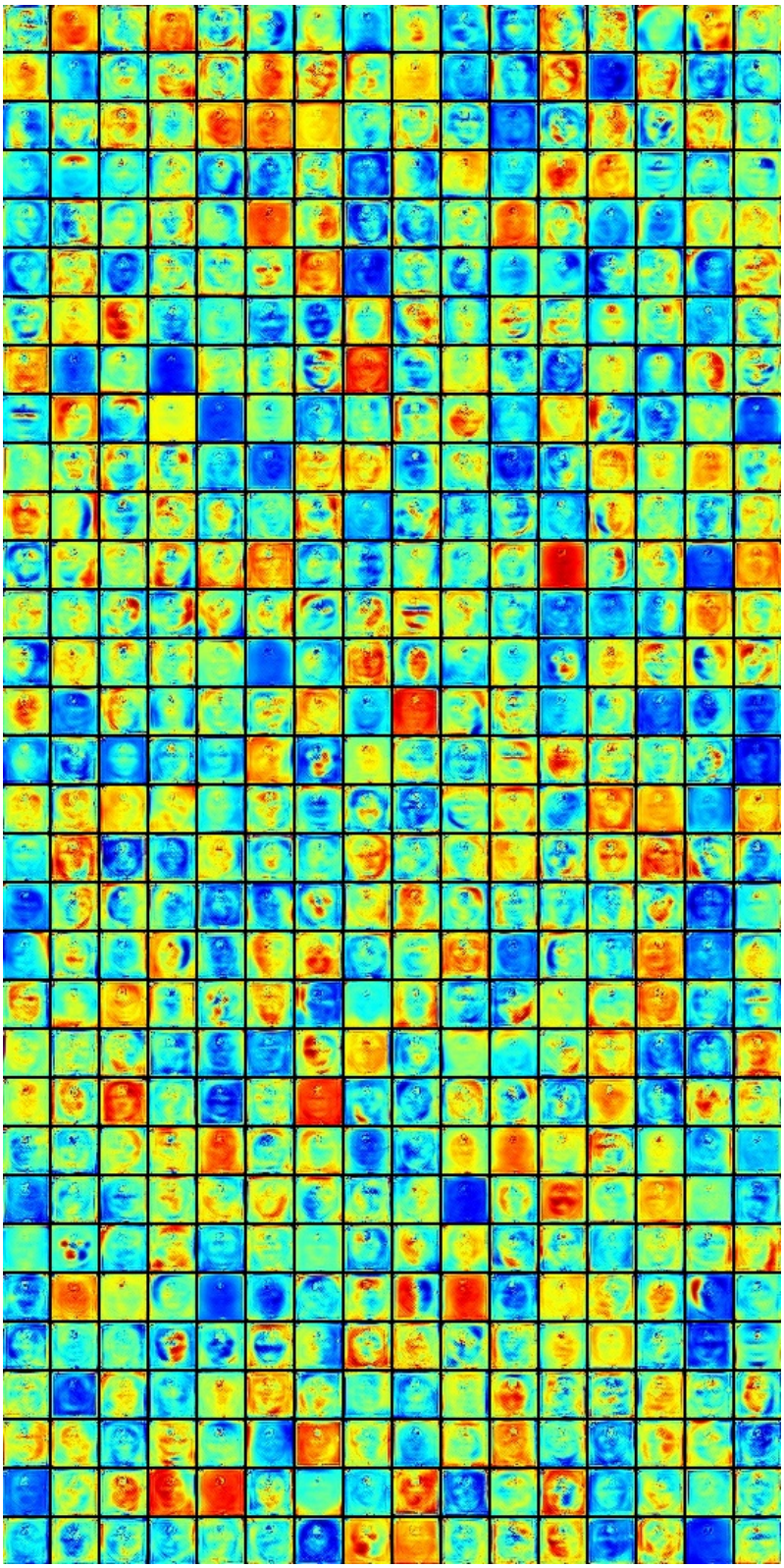

Figure 4: Visualization of all the channels of the meta memory $\mathbf{M}_o$.

Figure 5: Qualitative ablation studies in VoxCeleb1 dataset. The memory compensation module (MCM) and implicit identity representation conditioned memory module (IICM) can obtain improvements.

Figure 6: Qualitative ablation studies in VoxCeleb1 dataset. The memory compensation module (MCM) and implicit identity representation conditioned memory module (IICM) can obtain improvements.