LEARNING IMPLICIT SCALE CONDITIONED MEMORY COMPENSATION FOR TALKING HEAD GENERATION

Anonymous authors
Paper under double-blind review

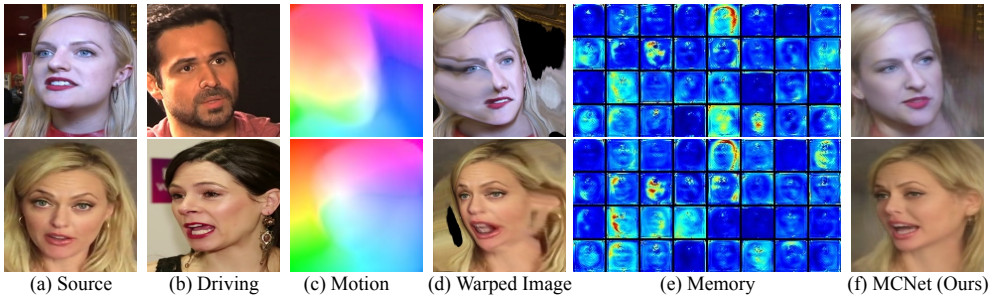
ABSTRACT

Talking head video generation aims to animate the pose and expression of a person in a target driving video using motion information contained in the video, while maintaining a person's identity in a given still source image. Highly dynamic and complex motions in the driving video cause ambiguous generation from the source image, because the still source image cannot provide sufficient appearance information for occluded regions or delicate expressions, which severely produces artifacts and significantly degrades the generation quality. However, existing works mainly focus on learning more accurate motion estimation and representation in 2D and 3D, and they ignore the facial structural prior in addressing the facial ambiguities. Therefore, effective handling of the ambiguities in the dramatic appearance changes of the source to largely improve facial details and completeness in generation still remains barely explored. To this end, we propose a novel implicit scale conditioned memory compensation network (MCNet) for high-fidelity talking head generation. Specifically, considering human faces are symmetric and structured, we aim to automatically learn a representative global facial memory bank from all training data as a prior to compensate the facial generation features. Each face in the source image contains a scale that can be reflected in detected facial keypoints. To better query the learned global memory, we further propose to learn implicit scale representations from the discrete keypoints, which can be used to condition on the query of the global memory, to obtain scale-aware memory for the feature compensation. Extensive experiments from quantitative and qualitative perspectives demonstrate that MCNet can learn representative and complementary facial memory, and can clearly outperform previous state-of-the-art methods on VoxCeleb1 and CelebV datasets.

1 INTRODUCTION

In this work, we aim to address the problem of generating a realistic talking head video given one still source image and one dynamic driving video, which is widely known as talking head video generation. A high-quality talking head generation model needs to imitate vivid facial expressions and complex head movements, and should be applicable for different facial identities presenting in the source image and the target video. It has been attracting rapidly increasing attention from the community, and a wide range of realistic applications remarkably benefits from this task, such as digital human broadcast, AI-based human conversation, and virtual anchors in films.

Significant progress has been achieved on this task in terms of both quality and robustness in recent years. Existing works mainly focus on learning more accurate motion estimation and representation in 2D and 3D to improve the generation. More specifically, 2D facial keypoints or landmarks are learned to model the motion flow (see Fig. 1(c)) between the source image and any target image in the driving video (Zhao et al. (2021); Zakharov et al. (2019); Hong et al. (2022)). Some works also consider utilizing 3D facial prior model (*e.g.* 3DMMBlanz & Vetter (1999)) with decoupled expression codes (Zhao et al., 2021; Zakharov et al., 2019) or learning dense facial geometries in a self-supervised manner (Hong et al., 2022) to model complex facial expression movements to produce more fine-grained facial generation. However, no matter how accurately the motion can be estimated and represented, highly dynamic and complex motions in the driving video cause ambiguous generation from the source image (see Fig. 1(d)), because the still source image cannot provide sufficient appearance information for occluded regions or delicate expressions, which severely produces artifacts and significantly degrades the generation quality.



(a) Source (b) Driving (c) Motion (d) Warped Image (e) Memory (f) MCNet (Ours)

Figure 1: MCNet animation illustration. MCNet first learns the motion flow (c) between the source and the driving images. (d) provides an elaboration of possible occulusion or deformation caused large motion, which is produced by directly warping the source image with the motion provided by the driving images. (e) shows randomly sampled memory channels of our learned scale-aware memory bank. We also present examples of generated results of our method in (f).

Intuitively, we understand that human faces are symmetrical and highly structured, and many regions of the human faces are essentially not discriminative. For instance, only blocking a very small eye region of a face image makes a well-trained facial recognition model largely drop the recognition performance (Qiu et al., 2021), which indicates to a certain extent that the structure and appearance representations of human faces crossing different face identities are generic and transferable. Therefore, learning global facial priors on spatial structure and appearance from all available training face images, and utilizing the learned facial priors for compensating the dynamic facial synthesis are critically important for high-fidelity talking head generation. However, existing works did not explore these beneficial facial priors to address facial ambiguities in generation from large head motions.

In this paper, to effectively deal with the ambiguities in the dramatic appearance changes from the still source image, we propose an implicit scale conditioned **M**emory **C**ompensation **Net**work, coined as **MCNet**, to learn and transfer global facial representations to compensate ambiguous facial details and guarantee completeness for a high-fidelity generation. Specifically, we design and learn a global spatial meta memory bank. The optimization gradients from all the training images during training contribute together to the updating of the meta memory, and thus it can capture the global and most common facial appearance and structure representations for the transferring. Since the different source face images contain distinct scales, to more effectively query the learned meta memory bank, we propose an implicit scale conditioned memory module (ISCM) (see Fig. 3). As the detected discrete facial keypoints inherently contain the scale information of the face, we first learn an implicit scale representation from the discrete keypoint coordinates, and further use it to condition on the query of the meta memory bank to obtain a scale-aware memory bank, which can more effectively compensate the feature of faces with different scales. The compensation is performed through a memory compensation module (MCM) (see Fig. 4). The warped feature map generated from the estimated motion field queries the scale-aware memory bank through a dynamic cross-attention mechanism to output a refined compensated feature map for the final generation.

We conduct extensive experiments to evaluate the proposed MCNet on two competitive talking head generation datasets (*i.e.* VoxCeleb (Nagrani et al., 2017) and CelebV (Wu et al., 2018). Experimental results demonstrate the effectiveness of learning global facial memory to tackle the appearance ambiguities in the talking head generation, and also show clearly improved generation results from both qualitative and quantitative perspectives, achieving state-of-the-art performances.

In summary, our main contribution is three-fold:

- We propose to learn a global facial meta memory bank to transfer representative facial representations to handle the appearance and structure ambiguities caused by the highly dynamic generation from a still source image. To the best of our knowledge, it is the first exploration in the literature to model global facial representations for effectively improving the ambiguities in talking head generation.
- We propose a novel implicit scale conditioned memory compensation network (MCNet) for talking head video generation, in which an implicit scale conditioned memory module (ISCM) and a facial memory compensation module (MCM) are designed to respectively perform the scale-aware memory learning and the feature compensation tasks.

- Qualitative and quantitative experiments extensively show the effectiveness of the learned meta memory bank for addressing the ambiguities in generation, and our framework establishes a clear state-of-the-art performance on the talking head generation. The generalization experiment also shows that the proposed modules can effectively boost the performance of different talking head generation models.

2 RELATED WORKS

Talking Head Video Generation. Talking Head video Generation can be mainly divided into two strategies: image-driven and audio-driven generation. For the image-driven strategy, researchers aim to capture the expression of a given driving image and aggregate the captured expression with the facial identity from a given source image. Some approaches (Yao et al., 2020; Wu et al., 2021b; Wang et al., 2021a) utilized a 3DMM regressor (Tran & Liu, 2018; Zhu et al., 2017) to extract an expression code and an identity code from a given face, and then respectively combine them from different faces to generate a new face. Also, several other works (Tripathy et al., 2021; Ha et al., 2020; Zakharov et al., 2020; 2019; Zhao et al., 2021) utilized facial landmarks detected by a pretrained face model (Guo et al., 2019) to act as anchors of the face. Then, the facial motion flow calculated from landmarks is transferred from a driving face video. However, their motion flow suffers from error accumulation caused by inaccuracy of the pretrained model. To overcome this limitation, the keypoints are learned in an unsupervised fashion (Siarohin et al., 2019; Hong et al., 2022; Wang et al., 2021b; Liu et al., 2021a; Zhao & Zhang, 2022) to better represent the motion of the face with carefully designed mechanisms for modeling the motion transformations between two sets of keypoints. Audio-driven talking head generation (Ji et al., 2022; Lu et al., 2021; Wu et al., 2021a; Ji et al., 2021) is another popular direction on this topic, as audio sequences do not contain information of the face identity, and is relatively easier to disentangle the motion information from the input audio. Liang et al. (2022) explicitly divides the driving audio into granular parts through delicate prior-based pre-processing to control the lip shape, face pose, and the facial expression.

In this work, we focus on the image-driven talking head generation. In contrast to previous image-driven works, we aim at learning global facial structure and appearance priors through a well-designed memory-bank network to effectively compensate intermediate facial features, which can produce higher-quality generation on ambiguous regions caused by large head motion.

Memory Bank Learning. Introducing an external memory / prior component is popular because of its flexible capability of storing, abstracting and organising long-term knowledge into a structural form. Recently, the memory bank has shown its powerful capabilities in learning and reasoning for addressing several challenging tasks, *e.g.* image processing (Yoo et al., 2019; Huang et al., 2021), video object detection (Sun et al., 2021), and image caption (Fei, 2021). As an earlier work, Weston et al. (2014) propose a memory network, which integrates inference components within a memory bank that can be read and written to memorize supporting facts from the past for question answering. Yoo et al. (2019) propose a memory-augmented colorization network to produce high-quality colorization with limited training data. Xu et al. (2021) uses the texture memory of patch samples extracted from unmasked regions to inpaint missing facial parts. Wu et al. (2022) proposes a memory-disentangled refinement network for coordinated face inpainting in a coarse-to-fine manner.

In contrast to these previous works, to the best of our knowledge, we are the first to propose using a global memory mechanism to deal with ambiguous generation issues in the task of talking head video generation. We also accordingly design a novel implicit-facial-scale-aware memory learning network and a novel memory compensation network to successfully tackle the issues.

3 METHODOLOGY

Under the same pipeline in previous work (Siarohin et al., 2019), we introduce an implicit scale conditioned memory compensation network, termed as MCNet, for talking head video generation. MCNet learns a facial-scale-aware memory bank by the designed implicit scale conditioned memory module (**ISCM**) to compensate the warped feature in the memory compensation module (**MCM**).

3.1 OVERVIEW

The framework of our MCNet depicted in Fig. 2 can be divided into three parts: (i) The keypoint detector and the dense motion network. Initially, the keypoint detector receives a source image $\mathbf{S}$ and a driving frame $\mathbf{D}$ to predict K pairs of keypoints, *i.e.* $\{x_{s,t}, y_{s,t}\}_{t=1}^{K}$ and $\{x_{d,t}, y_{d,t}\}_{t=1}^{K}$ on the source and target, respectively. With the keypoints generated from the driving frame and the

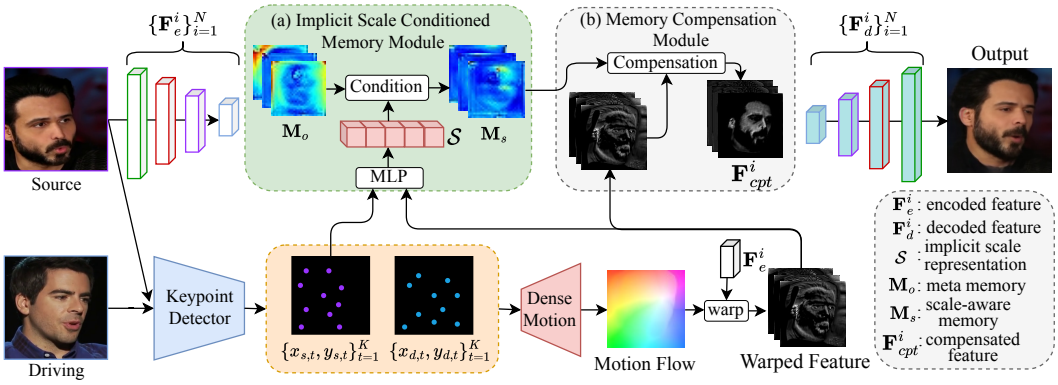


Figure 2: An illustration of the proposed MCNet, which contains two designed modules to compensate the facial feature map: (i) The implicit scale conditioned memory module (**ISCM**) learns the scale information from the input source, utilizing the warped feature map and keypoints of the source image, to produce an implicit scale representation $\mathcal{S}$, which is conditioned on the meta memory bank $\mathbf{M}_o$ to learn a scale-aware memory $\mathbf{M}_s$. (ii) The memory compensation module (**MCM**) adopts a dynamic cross-attention mechanism to compensate the warped feature map spatially.

source image, the dense motion network estimates the motion flow $A_{\mathbf{S}\leftarrow\mathbf{D}}$ between these two; (ii) The designed implicit scale conditioned memory module (ISCM). We first leverage the estimated motion flow $A_{\mathbf{S}\leftarrow\mathbf{D}}$ to warp the encoded feature $\mathbf{F}^i_e$ in the i-th layer, resulting in a warped feature $\mathbf{F}^i_w$. The warped feature $\mathbf{F}^i_w$ and the source keypoints are fed into the implicit scale conditioned memory module to encode an implicit scale representation, which will be conditioned on the query of the meta memory $\mathbf{M}_o$ to produce an identity-dependent scale-aware memory bank $\mathbf{M}_s$; (iii) The memory compensation module (MCM). After obtaining the scale-aware memory bank $\mathbf{M}_s$, we utilize a dynamic cross-attention mechanism to compensate the warped features spatially in the MCM, and then output a compensated feature $\mathbf{F}^i_{cpt}$. Finally, our decoder utilizes all the N feature maps *i.e.* $\{\mathbf{F}^i_{cpt}\}^N_{i=1}$, to produce the final image $\mathbf{I}_{rst}$. In the following, we will show how to learn our memory bank in the ISCM and how it is utilized in the MCM for generation-feature compensation.

3.2 Learning Implicit-Scale-Conditioned Global Facial Memory

We first aim at learning a global meta-memory bank to model facial structure and appearance representations from the whole face dataset. It is clear that human faces have multiple scales in the real-world, thus learning a meta-memory bank to directly compensate all source faces with different scales is inflexible. To handle the faces with distinct scales, we design an implicit scale conditioned memory module (ISCM) to learn a scale-aware memory, through source-scale-conditioned query on the global meta-memory bank to compensate warped source face features with scale variations.

Meta memory. In this work, we first aim to learn a global meta memory bank to store the global and generic facial appearance and spatial structure representations from all the training data available. We initialize a meta memory bank $\mathbf{M}_o$ as a cube tensor with a shape of $C_m \times H_m \times W_m$ instead of a vector (Esser et al., 2021). Moreover, the multiple channels hold enough capacity for the meta memory to learn different facial structures and appearances (see Fig. 7). As many regions of the human faces are not discriminative and transferable, we can utilize the global facial priors learned in the meta memory to compensate ambiguous regions in the generated faces. With a designed objective function, the meta memory bank can be automatically updated by the optimization gradients from all the training images during the training stage. In this way, the facial prior learned in the meta memory is global rather than conditioned on any specific input sample, which provides highly beneficial global information for face compensation in generation.

Implicit scale representation learning. In our framework, the detected facial keypoints are used to learn motion flow for feature warping. The facial keypoints implicitly contain the scale information of the human face because of their structural positions (Siarohin et al., 2019; Tao et al., 2022). Therefore, we utilize both the source keypoints $\{x_{s,t}, y_{s,t}\}^K_{t=1}$ and the warped feature $\mathbf{F}^i_w$ to learn an implicit scale representation of the source face. The reason of learning a scale representation for the source is that we aim to compensate the image with the identity of the source image. And the warped feature $\mathbf{F}^i_w$ is used because it also contains the scale information of the source image, as the warped feature is generated via warping the source feature with the keypoint-based motion flow.

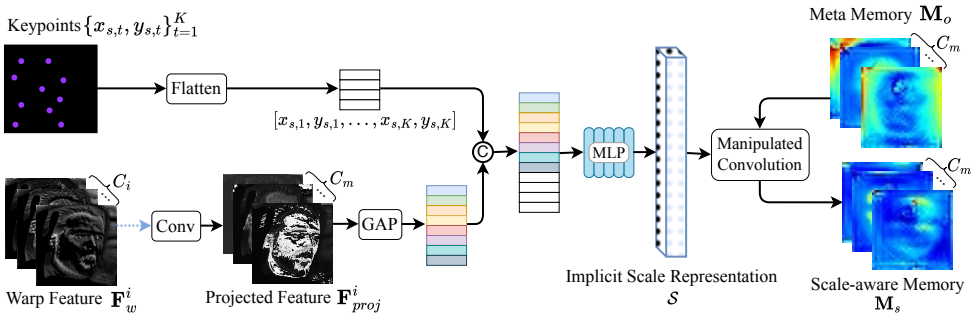


Figure 3: The illustration of the proposed implicit scale conditioned memory module (ISCM). The symbol ⓒ denotes the concatenation operation, and the "GAP" and "Conv" represent the global average pooling and the convolution layer, respectively. The detailed generation of the projected feature $\mathbf{F}_{proj}^i$ can refer to Fig. 4. C_i denotes the channel number of the i-th level warped feature $\mathbf{F}_w^i$ in our autoencoder framework, while C_m is the channel number of our memory bank.

As shown in Fig. 3, to embed the scale information, we first utilize a global average pooling function $\mathcal{F}_{GAP}$ to squeeze the global spatial information of the projected feature $\mathbf{F}_{proj}^i$, which is produced from the warped feature $\mathbf{F}_w^i$ (see Fig. 4), into a channel descriptor. After that, we concatenate the flattened and normalized keypoints with the feature vector from $\mathcal{F}_{GAP}$, and feed them into an MLP mapping network $\mathcal{F}_{mlp}$ to learn an implicit scale representation $\mathcal{S}$ of the source image. Thus:

$$\mathcal{S} = \mathcal{F}_{mlp}(\mathcal{S}'), \quad \mathcal{S}' = \text{Concat}\left[\mathcal{F}_{GAP}(\mathbf{F}_{proj}^i), [x_{s,1}, y_{s,1}, \ldots, x_{s,K}, y_{s,K}]\right], \tag{1}$$

where $\text{Concat}[\cdot,\cdot]$ denotes the concatenation operation. In this way, the implicit scale representation $\mathcal{S}$ of the source image can be learned.

Scale-aware memory learning. As discussed before, human faces present a diverse range of scales in reality. Compared to directly using the global meta memory for facial-feature compensation, we believe that a scale-dependent condition on the meta-memory is a more intuitive and effective way. Therefore, we propose to condition the learned implicit scale representation $\mathcal{S}$ on the meta memory $\mathbf{M}_o$ to obtain an identity-dependent scale-aware memory $\mathbf{M}_s$ for each face image. Inspired by the style injection in StyleGANv2 (Karras et al., 2020), we utilize the implicit scale representation $\mathcal{S}$ to manipulate a 3×3 convolution layer to produce the implicit scale-aware facial memory: $\omega'_{ijk} = s_i * \omega_{ijk}$ and $\omega''_{ijk} = \omega'_{ijk} \sqrt{\sum_{i,k} (\omega'_{ijk})^2 + \epsilon}$, where ω is the weight of the convolution kernel, ϵ is a small constant to avoid numerical issues, s_i is the i-th element in the learned implicit scale representation $\mathcal{S}$, and j and k enumerate the output feature maps and spatial footprint of the convolution, respectively. Finally, we obtain the scale-aware memory as:

$$\mathbf{M}_s = \mathcal{F}_{C_{\omega''}}(\mathbf{M}_o) \tag{2}$$

where the $\mathcal{F}_{C_{\omega''}}$ is the manipulated convolution layer parameterized by ω''. With the scale-aware memory bank $\mathbf{M}_s$, each input sample can be compensated by the scale-correlated facial priors, resulting in better generation performance discussed in the experiments.

3.3 Global Memory Compensation and Generation

The warped feature map contains ambiguity for generation caused by large head motion or occlusion, we thus propose to inpaint those ambiguous regions via compensating the warped facial features. To this end, we design a memory compensation module (MCM, see Fig. 4) to refine the warped feature $\mathbf{F}_w^i$ via the learned scale-aware facial memory bank.

Warped facial feature projection. To maintain better the identity information in the source image while compensating the warped feature from the source, we employ a channel-split strategy to split the warped feature $\mathbf{F}_w^i$ into two parts along the channel dimension: $\mathbf{F}_w^{i,0}$ and $\mathbf{F}_w^{i,1}$. The part of the first half channels $\mathbf{F}_w^{i,0}$ is left to directly pass through for contributing for the identity preserving, while the part of the rest half channels $\mathbf{F}_w^{i,1}$ is modulated by the learned scale-aware memory bank $\mathbf{M}_s$ to refine the ambiguities. After splitting, we employ a 1×1 convolution layer on $\mathbf{F}_w^{i,1}$ to change the channel number, resulting in a projected feature $\mathbf{F}_{proj}^i$.

Warped facial feature compensation. To compensate the feature map spatially, we adopt a dynamic cross-attention mechanism to refine features. Specifically, we employ the scale-aware mem-

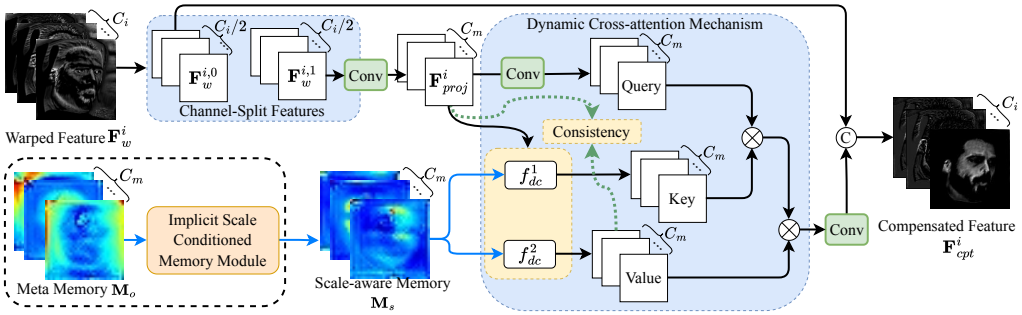


Figure 4: The illustration of the memory compensation module (MCM). The symbol $\bigotimes$ denotes matrix multiplication, and f_{dc}^1 and f_{dc}^2 are dynamic convolution layers (Chen et al., 2020), whose kernel weights are estimated by the projected feature $\mathbf{F}_{proj}^i$. The © represents the concatenation operation, and the "Conv" denote a convolution layer. C_i is the channel number of the i-th level feature in our autoencoder framework, while C_m is the channel number of the memory bank.

ory to produce the Key $\mathbf{F}_K^i$ and Value $\mathbf{F}_V^i$ via two dynamic convolution layers (*i.e.* f_{dc}^1, f_{dc}^2) conditioned on the projected feature $\mathbf{F}_{proj}^i$. In this way, the generated Key and Value are identity-dependent and capable of providing useful context information. We, in the meanwhile, perform a non-linear projection to map $\mathbf{F}_{proj}^i$ into a query feature $\mathbf{F}_Q^i$ by a 1×1 convolution layer followed by a ReLU layer. Then, we perform cross attention to reconstruct a more robust feature $\mathbf{F}_{ca}^i$ as:

$$\mathbf{F}_{ca}^i = \mathcal{F}_{C_{1\times1}}\left(\text{Softmax}\left({\mathbf{F}_Q^i}^T \times \mathbf{F}_K^i\right) \times \mathbf{F}_V^i\right) \tag{3}$$

where "Softmax" denotes the softmax operator, while the $\mathcal{F}_{C_{1\times1}}$ is a 1×1 convolution layer to change the channel number of the cross-attention output. "$\times$" denotes a matrix multiplication. As shown in Fig. 4, to maintain the identity of the source image, we concatenate the cross-attention features $\mathbf{F}_{ca}^i$ with the first half-channels $\mathbf{F}_w^{i,0}$:

$$\mathbf{F}_{cpt}^i = \text{Concat}[\mathbf{F}_{ca}^i, \mathbf{F}_w^{i,0}], \tag{4}$$

where the $\text{Concat}[\cdot,\cdot]$ represents a concatenation operation. As a result, the final output feature $\mathbf{F}_{cpt}^i$ enjoys the benefits of directly incorporating learned facial prior information (Wang et al., 2021c) from the memory and effectively modulating by the dynamic cross-attention mechanism.

Regularization on consistency. To learn the global and the most generic spatial facial appearance and structure representations from the input faces, we need to make the learning of the meta memory constrained by each single image in the training data. Simply but effectively, we enforce the consistency between the projected feature $\mathbf{F}_{proj}^i$ from the current training face image, and the value feature $\mathbf{F}_V^i$ from the global meta memory:

$$\mathcal{L}_{con} = ||\mathbf{F}_V^i - de(\mathbf{F}_{proj}^i)||_1, \tag{5}$$

where the $de(\cdot)$ indicates a gradient detach function and $||\cdot||_1$ is $\mathcal{L}_1$ loss. By using this function, the regularization enforces the consistency on the learning of the global meta-memory while not affecting the learning of the source image features to guarantee the training stability of the overall generation framework. The above equation also makes sure that the optimization gradients from all the face images during the training state contribute together to the updating of the memory bank, and thus it can capture global facial appearance and structure representations for the transferring.

Multi-layer generation. A higher-resolution feature map contains more facial details, while a smaller-resolution one contains more semantic information. We perform memory compensation for feature maps of each layer to preserve facial details as TPSM (Zhao & Zhang, 2022). As shown in Fig. 2, We utilize the motion flow $A_{\mathbf{S}\leftarrow\mathbf{D}}$ to warp the encoded feature $\{\mathbf{F}_e^i\}_{i=1}^N$ in each layer to produce warped features $\{\mathbf{F}_w^i\}_{i=1}^N$. For each warped feature $\mathbf{F}_w^i$, we feed it into our designed ISCM and the MCM modules sequentially to produce the compensated features $\{\mathbf{F}_{cpt}^i\}_{i=1}^N$. In the decoding process, we treat the $\mathbf{F}_{cpt}^1$ as $\mathbf{F}_d^1$ and then the $\mathbf{F}_d^2$ is generated by $\mathbf{F}_d^1$ through an upsampling layer. At i-th level ($i > 1$), the output compensated feature $\mathbf{F}_{cpt}^i$ will be concatenated with the decoded feature $\mathbf{F}_d^i$ at the same level to produce a decoded feature $\mathbf{F}_d^{i+1}$ at the next level via an upsampling layer. Finally, we input the concatenation of $\mathbf{F}_d^N$ and $\mathbf{F}_{cpt}^N$ into a convolution layer followed by a Sigmoid unit to generate the final facial image $\mathbf{I}_{rst}$. Each layer shares the same meta memory $\mathbf{M}_o$.

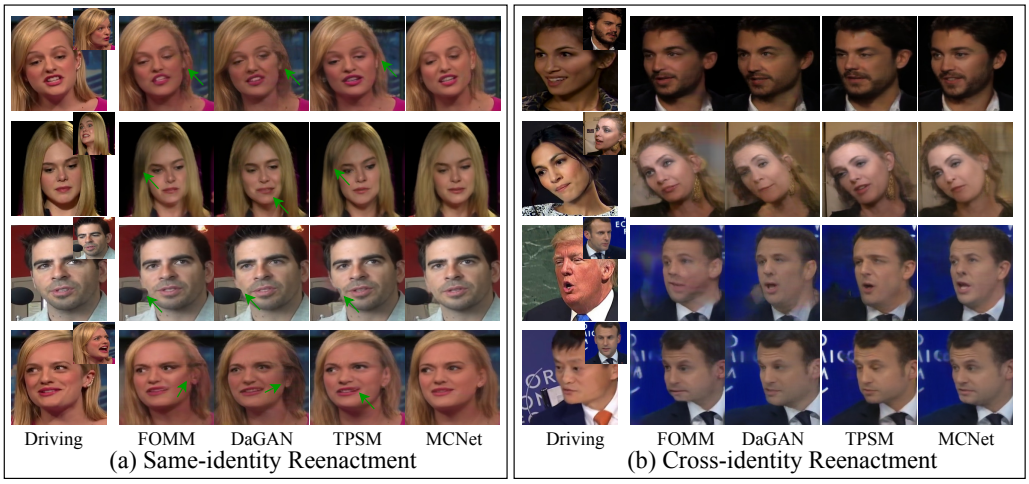


Figure 5: Qualitative comparisons of (a) same-identity reenactment and (b) cross-identity reenactment on the VoxCeleb1 (the first two rows) and CelebV dataset (the last two rows). Our method shows higher-fidelity generation compared to the state-of-the-arts. **Zoom in for best view.**

3.4 Training

Loss objectives. We train the proposed MCNet by minimizing the following losses:

$$\mathcal{L} = \lambda_P \mathcal{L}_P + \lambda_{eq} \mathcal{L}_{eq} + \lambda_{dist} \mathcal{L}_{dist} + \lambda_{con} \mathcal{L}_{con} \tag{6}$$

where the λ_P, λ_{eq}, λ_{dist} and λ_{con} are the hyper-parameters to allow for a balanced learning from these losses. Per FOMM (Siarohin et al., 2019), we leverage the perceptual loss $\mathcal{L}_P$ to minimize the gap between the model output and the driving image and equivariance loss $\mathcal{L}_{eq}$ to learn more stable keypoints. Additionally, we also adopt the keypoints distance loss $\mathcal{L}_{dist}$ (Hong et al., 2022) to avoid the detected keypoints crowding around a small neighbourhood. The $\mathcal{L}_{con}$ is the consistency loss in Eq. 5. The details of these losses are described in Appendix.

4 Experiments

In this section, we present quantitative and qualitative experiments to validate the performance of our MCNet. More details (*i.e.* additional results and implementation) are included in the Appendix.

4.1 Datasets and Metrics

Dataset. In this work, we mainly evaluate our MCNet on two talking head generation datasets, *i.e.* VoxCeleb1 (Nagrani et al., 2017) and CelebV (Wu et al., 2018) dataset. We follow the sampling strategy for the test set in DaGAN (Hong et al., 2022) for evaluation. Following DaGAN, to verify the generalization ability, we apply the model trained on VoxCeleb1 to test on CelebV.

Metrics. We adopt the structured similarity (**SSIM**), peak signal-to-noise ratio (**PSNR**), and $\mathcal{L}_1$ distance to measure the low-level similarity between the generated image and the driving image. Following the previous works (Siarohin et al., 2019), we utilize the Average Euclidean Distance (**AED**) to measure the identity preservation, and Average Keypoint Distance (**AKD**) to evaluate whether the motion of the input driving image is preserved. We also adopt the **AUCON** and **PRMSE**, similar to Hong et al. (2022), to evaluate the expression and head poses in cross-identity reenactment.

4.2 Comparison with state-of-the-art methods

Same-identity reenactment. In Table 1(a), we first compare the synthesised results for the setup in which the source and the driving images share the same identity. It can be observed that our MCNet obtains the best results compared with other competitive methods. Specifically, compared with FOMM (Siarohin et al., 2019) and DaGAN (Hong et al., 2022), which adopt the same motion estimation method as ours, our method can produce higher-quality images (72.3% of FOMM vs 82.5% of ours, resulting in a 10.2% improvement on the SSIM metric), which verifies that introducing the global memory mechanism can indeed benefit the image quality in the generation process. Regarding motion animation and identity preservation, our MCNet also achieves the best results (*i.e.* 1.203 on AKD and 0.106 on AED), showing superior performance on the talking head animation.

Figure 6: Qualitative ablation studies. The memory compensation module (MCM) and implicit scale conditioned memory module (ISCM) can obtain improvements. The last column verifies that our ISCM can learn different scale-aware memories for different scale samples.

Table 1: Comparisons with state-of-the-art methods on (a) same-identity reenactment on VoxCeleb1 and (b) cross-identity reenactment on VoxCeleb1 and CelebV dataset.

Model	(a) Results of Same-identity Reenactment						(b) Results of Cross-identity Reenactment			
	VoxCeleb1						VoxCeleb1		CelebV1	
	SSIM (%) ↑	PSNR ↑	LPIPS ↓	$\mathcal{L}_1$ ↓	AKD ↓	AED ↓	AUCON ↑	PRMSE ↓	AUCON ↑	PRMSE ↓
X2face (Wiles et al., 2018)	71.9	22.54	-	0.0780	7.687	0.405	-	-	0.679	3.62
marioNETte (Ha et al., 2020)	75.5	23.24	-	-	-	-	-	-	0.710	3.41
FOMM (Siarohin et al., 2019))	72.3	30.39	0.199	0.0430	1.294	0.140	0.882	2.824	0.667	3.90
MeshG (Yao et al., 2020)	73.9	30.39	-	-	-	-	-	-	0.709	3.41
face-vid2vid (Wang et al., 2021b)	76.1	30.69	0.212	0.0430	1.620	0.153	0.839	4.398	0.805	3.15
MRAA (Siarohin et al., 2021)	80.0	31.39	0.195	0.0375	1.296	0.125	0.882	2.751	0.840	2.46
DaGAN (Hong et al., 2022)	80.4	31.22	0.185	0.0360	1.279	0.117	0.888	2.822	0.873	2.33
TPSN (Zhao & Zhang, 2022)	81.6	31.43	0.179	0.0365	1.233	0.119	0.894	2.756	0.882	2.23
MCNet (Ours)	**82.5**	**31.94**	**0.174**	**0.0331**	**1.203**	**0.106**	**0.895**	**2.641**	**0.885**	**2.10**

Moreover, we show several samples in Fig. 5(a), and the face samples in Fig. 5(a) contain large motions (the first, the third, and the last row) and object occlusion (the second row). From Fig. 5(a), our model can effectively handle these complex cases and produces more completed image generations compared with the state-of-the-art competitors.

Cross-identity reenactment. We also perform experiments on the VoxCeleb1 and CelebV datasets to conduct the task of the cross-identity face motion animation, in which the source and driving images are from different people. The results compared with other methods are reported in Table 1. Our MCNet outperforms all the other comparison methods. Regarding the head pose imitation, our MCNet can produce the face with a more accurate head pose (*i.e.* 2.641 and 2.10 for VoxCeleb1 and CelebV, respectively, on the PRMSE metric). We also present several samples of results with the VoxCeleb1 dataset in Fig. 5(b). It is clear to observe that our MCNet can mimic the facial expression better than the other methods, such as the smiling countenance shown in the first row. For the unseen person in the CelebV dataset, *e.g.* the last two rows in Fig. 5(b), our method can still produce a more natural generation, while the results of other methods contain more obvious artifacts. All of these results verify that the feature compensated by our learned memory can produce better results.

4.3 Ablation Study

In this section, we perform ablation studies to demonstrate the effectiveness of the proposed implicit scale conditioned memory module (ISCM) and memory compensation module (MCM). We report the quantitative results in Table 2 and the qualitative results in Fig. 6. Our baseline is the model without ISCM and MCM modules. The "Baseline + MCM" means that we drop the ISCM module and replace the scale-aware memory $\mathbf{M}_s$ with the meta memory $\mathbf{M}_o$ in Fig. 4.

Meta Memory learning. We first visualize the learned meta memory in Fig. 7, which aims to learn the global and generic facial appearance and structure representations. In Fig. 7, we visualize partial channels of the meta memory bank. It can be observed that all these channels represent the faces with different appearances, structures, poses, and shapes, which are very informative and clearly beneficial for the facial compensation and generation, confirming our motivation of learning global facial representations to tackle ambiguities in the talking head generation.

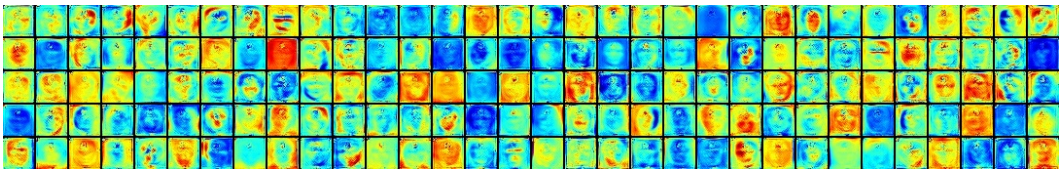


Figure 7: The visualization of randomly selected channels of the meta memory $\mathbf{M}_o$. We can observe that our meta memory learns very diverse global facial representations. **Zoom in for best view.**

Table 2: Ablation studies: "Baseline" indicates the simplest model without the implicit scale conditioned memory module (ISCM) and memory compensation module (MCM). "$\mathrm{MCM}^{\mathrm{w/o\,Eq.4}}$" means that we project the entire warped feature into a projected feature $\mathbf{F}^i_{proj}$, and remove the concatenation function Eq.5 to make the output of cross-attention as the compensated feature $\mathbf{F}^i_{cpt}$.

Model	SSIM (%) ↑	PSNR ↑	LPIPS ↓	$\mathcal{L}_1$ ↓	AKD ↓	AED ↓
Baseline	81.1	31.70	0.182	0.0356	1.303	0.124
Baseline + $\mathrm{MCM}^{\mathrm{w/o\,Eq.4}}$	82.0	31.82	0.176	0.0340	1.242	0.119
Baseline + MCM	82.3	31.92	0.175	0.0334	1.237	0.114
Baseline + ISCM + MCM (MCNet)	**82.5**	**31.94**	**0.174**	**0.0331**	**1.203**	**0.106**
FOMM (Siarohin et al., 2019)	72.3	30.39	0.199	0.0430	1.294	0.140
FOMM+ ISCM + MCM	81.8	31.73	0.179	0.0353	1.269	0.119
TPSN (Zhao & Zhang, 2022)	81.6	31.43	0.179	0.0365	1.233	0.119
TPSN+ ISCM + MCM	82.0	31.55	0.175	0.0356	1.216	0.115

Memory compensation In Table 2 and Fig. 6, the proposed memory compensation module can effectively improve the generation quality of human faces. From Tab. 2, we observe that adding the memory compensation module (MCM) can consistently boost the performance via comparison between "Baseline+MCM" and "Baseline" (82.3% vs. 81.1% on SSIM). In Fig. 6, we also can see that the variant "Baseline+MCM" compensates the warped image better than the "Baseline", *e.g.* the face shape in the second row and the mouth shape in the third row. Additionally, we also conduct an ablation study to verify the feature channel split strategy discussed in Sec. 3.3. The results of "Baseline + $\mathrm{MCM}^{\mathrm{w/o\,Eq.4}}$" show that the channel split can slightly improve the performance. All these results demonstrate that learning a global facial memory can indeed effectively compensate the warped facial feature to produce higher-fidelity results for the talking head generation.

Scale-aware memory learning. To verify the effectiveness of the implicit scale conditioned memory module (*i.e.* ISCM introduced in Sec. 3.2), we show the randomly sampled channels of the scale-aware memory in Fig 6. As shown in the last column in Fig 6, the ISCM can produce an identity-dependent scale-aware memory bank, which have structural and scale relations with the input source images. By deploying the ISCM, our MCNet can produce highly realistic-looking images compared with the "Baseline+MCM", verifying that the learned scale-aware memory condioned on the input source can provide a better compensatation on the source feature for more vivid generation.

Generalization experiment. Importantly, we also insert the proposed MCM and ISCM modules into FOMM (Siarohin et al., 2019) and TPSM (Zhao & Zhang, 2022) to verify our designed memory mechanism can be flexibly generalized to existing talking head models. As shown in Table 2, the TPSM, which has a different motion estimation method compared to ours, deployed with our proposed memory modules, can achieve a stable improvement. The "FOMM+ISCM+MCM" can also gain a significant improvement on SSIM compared with the pioneering work "FOMM". These results demonstrate the transferability and generalization capabilities of the proposed method.

5 Conclusion

In this paper, we present an implicit scale conditioned memory compensation network (MCNet) to learn a global facial prior of spatial structure and appearance to address the ambiguity problem caused by the dynamic motion in the talking head video generation task. MCNet utilizes a designed implicit scale conditioned memory module to learn the scale-aware memory for different samples, which will be used to compensate the feature map in the memory compensate module. Ablation studies clearly show the effectiveness of learning global facial meta memory for the talking head video generation task. Our MCNet also produces more natural-looking results compared with the state-of-the-art on all benchmarks.

REFERENCES

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

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999.

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.

Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *CVPR*, 2020.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

Zhengcong Fei. Memory-augmented image captioning. In *AAAI*, 2021.

Xiaojie Guo, Siyuan Li, Jinke Yu, Jiawan Zhang, Jiayi Ma, Lin Ma, Wei Liu, and Haibin Ling. Pfld: A practical facial landmark detector. *arXiv preprint arXiv:1902.10859*, 2019.

Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *AAAI*, 2020.

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

Huaibo Huang, Aijing Yu, and Ran He. Memory oriented transfer learning for semi-supervised image deraining. In *CVPR*, 2021.

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

Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. *arXiv preprint arXiv:2205.15278*, 2022.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020.

Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Expressive talking head generation with granular audio-visual control. In *CVPR*, 2022.

Peirong Liu, Rui Wang, Xuefei Cao, Yipin Zhou, Ashish Shah, Maxime Oquab, Camille Couprie, and Ser-Nam Lim. Self-appearance-aided differential evolution for motion transfer. *arXiv preprint arXiv:2110.04658*, 2021a.

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*, 2021b.

Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: real-time photorealistic talking-head animation. *TOG*, 2021.

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.

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.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

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

Haibo Qiu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. End2end occluded face recognition by masking corrupted features. *TPAMI*, 2021.

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

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

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

Guanxiong Sun, Yang Hua, Guosheng Hu, and Neil Robertson. Mamba: Multi-level aggregation via memory bank for video object detection. In *AAAI*, 2021.

Jiale Tao, Biao Wang, Borun Xu, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. Structure-aware motion transfer with deformable anchor model. In *CVPR*, 2022.

Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *CVPR*, 2018.

Soumya Tripathy, Juho Kannala, and Esa Rahtu. Facegan: Facial attribute controllable reenactment gan. In *WACV*, 2021.

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.

Qiulin Wang, Lu Zhang, and Bo Li. Safa: Structure aware face animation. In *3DV*, 2021a.

Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021b.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021c.

Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *ECCV*, 2018.

Haozhe Wu, Jia Jia, Haoyu Wang, Yishun Dou, Chao Duan, and Qingshan Deng. Imitating arbitrary talking style for realistic audio-driven talking face synthesis. In *ACM Multimedia*, 2021a.

Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *ECCV*, 2018.

Xintian Wu, Qihang Zhang, Yiming Wu, Huanyu Wang, Songyuan Li, Lingyun Sun, and Xi Li. F^3a-gan: Facial flow for face animation with generative adversarial networks. *TIP*, 2021b.

Zhuojie Wu, Xingqun Qi, Zijian Wang, Wanting Zhou, Kun Yuan, Muyi Sun, and Zhenan Sun. Showface: Coordinated face inpainting with memory-disentangled refinement networks. *arXiv preprint arXiv:2204.02824*, 2022.

Rui Xu, Minghao Guo, Jiaqi Wang, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Texture memory-augmented deep patch-based image inpainting. *IEEE Transactions on Image Processing*, 2021.

Guangming Yao, Yi Yuan, Tianjia Shao, and Kun Zhou. Mesh guided one-shot face reenactment using graph convolutional networks. In *ACM MM*, 2020.

Seungjoo Yoo, Hyojin Bahng, Sunghyo Chung, Junsoo Lee, Jaehyuk Chang, and Jaegul Choo. Coloring with limited data: Few-shot colorization via memory augmented networks. In *CVPR*, 2019.

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

Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *ECCV*, 2020.

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.

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

Ruiqi Zhao, Tianyi Wu, and Guodong Guo. Sparse to dense motion transfer for face image animation. In *ICCV*, 2021.

Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *TPAMI*, 2017.

APPENDIX

A REPRODUCIBILITY

In this work, we utilize the PyTorch framework to implement our method. We develop our code based on the FOMM[1]. All experiments in this work are conducted on publicly available datasets. Hyperparameters necessary for reproducing our experiments are reported in "Implementation Details".

B TRAINING DETAILS AND ADDITIONAL NETWORK

B.1 IMPLEMENTATION DETAILS

The keypoint estimator and dense motion are borrowed from the FOMM (Siarohin et al., 2019). 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 that the ground-truth is the driving frame during training. To optimize the training objectives, we set the $\lambda_{rec} = 10$, $\lambda_{eq} = 10$, $\lambda_{dist} = 10$ and $\lambda_{con} = 10$. The number of keypoints in this work is 15, which is the same as that of DaGAN (Hong et al., 2022). 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 totally. The number of layers of encoder and decoder N is set as 4 and the number of keypoints K is 15 as Hong et al. (2022). We set the size of meta memory as $512 \times 32 \times 32$, where $C_m = 512$, $H_m = 32$ and $W_m = 32$. In the warping process, for those $\mathbf{F}_e^i$ that has a different spatial size as the motion flow, we employ the bilinear interpolation method to adjust the spatial size of the motion flow.

B.2 LOSS DETAILS

Perceptual Loss $\mathcal{L}_p$. Perceptual loss is a popular objective function in image generation tasks. As introduced in DaGAN (Hong et al., 2022), a generated image and its ground-truth, *i.e.* 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 (Simonyan & Zisserman, 2014) to extract the features from each resolution image. 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 4 different resolutions of ground-truth. Then, we measure the $\mathcal{L}_1$ distance between the ground-truth and generated image as the Perceptual loss:

$$\mathcal{L}_p = \sum_{i=1}^{4} \mathcal{L}_1(G_i, R_i) \tag{7}$$

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 (Siarohin et al., 2019), given an image I and its detected keypoints $\{X_i\}_{i=1}^K$ ($X_i \in \mathbb{R}^{1\times 2}$), we then perform a known spatial transformation T on images I and keypoints $\{X_i\}_{i=1}^K$, resulting in transformed image I_T and transformed keypoints $\{X_i^T\}_{i=1}^K$. Then, we use detect the keypoints on the transformed image I_T, denoted as $(\{X_{I_T,i}\}_{i=1}^K)$. We employ the equivariance Loss on the source image and driving image:

$$\mathcal{L}_{eq} = \sum_{i=1}^{K} ||X_i^T - X_{I_T,i}||_1 \tag{8}$$

Keypoint distance loss $\mathcal{L}_{dist}$. We employ the keypoint distance loss as Hong et al. (2022) 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 have:

[1] https://github.com/AliaksandrSiarohin/first-order-model

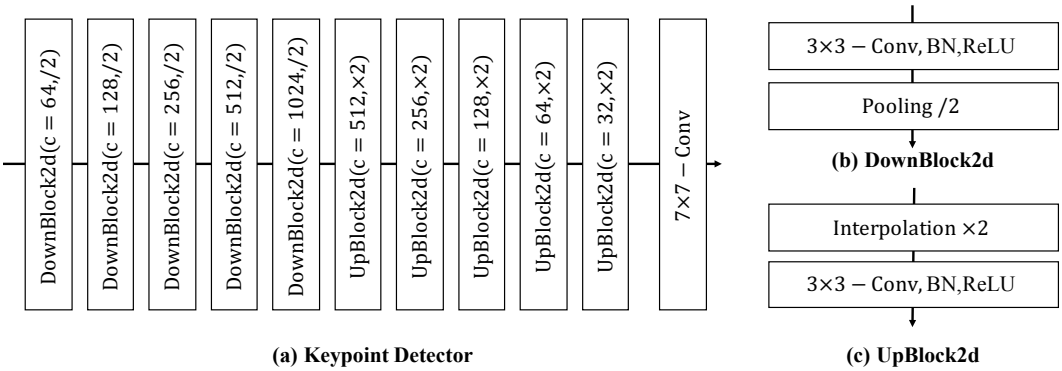


Figure 8: The structure of keypoint detecotor. The c in each layer means the output channel.

$$\mathcal{L}_{dist} = \sum_{i=1}^{K}\sum_{j=1}^{K}(1 - \mathbf{sign}||X_i - X_j||_1 - \alpha), i \neq j, \tag{9}$$

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

B.3 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 keypoint detector is illustrated in Fig. 8. Here, we adopt the Taylor approximation as FOMM (Siarohin et al., 2019) and DaGAN (Hong et al., 2022) to compute the motion flow. Thus, the motion estimation is not our contribution and we mainly focus on our meta memory.

C Experiments

C.1 Metrics

We mainly introduce four important metrics in talking head talks 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 (Baltrušaitis et al., 2016) to extract identity embeddings from the reconstructed face and the ground truth frame pairs. The MSE loss is employed to measure their difference.

Average keypoint distance (ADK). ADK evaluates the difference between landmarks of the reconstructed faces and ground truth frames. We extract the facial landmark using the face alignment method (Bulat & Tzimiropoulos, 2017). We compute the average distance between corresponding keypoints. Thus, the AKD mainly measures the ability of pose imitation.

The root mean square error of the head pose angles (PRMSE). In this work, we utilize the Py-Feat toolkit[2] to detect the Euler angles of 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.

C.2 Other experimental results

Positional Encoding for keypoints. The positional encoding method shows its strong power in the transformer (Vaswani et al., 2017; Liu et al., 2021b; Dosovitskiy et al., 2020) and Nerf (Mildenhall

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

Table 3: The results of applying positional encoding function on keypoints. "pe(10)" means we set the output dimension control factor $\mathbf{L}$ of positional encoding function as 10, and 20 for "pe(20)"

Model	SSIM (%) ↑	PSNR ↑	LPIPS ↓	$\mathcal{L}_1$ ↓	AKD ↓	AED ↓
Ous w/ pe(10)	82.4	31.91	0.175	0.0334	1.221	0.1073
Ous w/ pe(20)	69.4	30.03	0.269	0.0593	5.544	0.2684
MCNet	**82.5**	**31.94**	**0.174**	**0.0331**	**1.203**	**0.1065**

Table 4: Ablation studies. 'ISCM w/o $\mathbf{F}^i_{proj}$" and "ISCM w/o keypoints" represent that ISCM does not use the projected feature $\mathbf{F}^i_{proj}$ or keypoints as input (see Fig. 3), respectively, to encode implicit scale representation. "MCM 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 Key and Value in MCM.

Model	SSIM (%) ↑	PSNR ↑	LPIPS ↓	$\mathcal{L}_1$ ↓	AKD ↓	AED ↓
MCNet (ISCM w/o $\mathbf{F}^i_{proj}$)	82.3	31.89	0.175	0.0336	1.237	0.1097
MCNet (ISCM w/o keypoints)	82.4	31.93	0.175	0.0333	1.227	0.1087
MCNet (MCM w/o f^1_{dc}, f^2_{dc})	82.2	31.89	0.176	0.0336	1.246	0.1119
MCNet (single layer)	82.3	31.90	0.175	0.0334	1.235	0.1085
MCNet	**82.5**	**31.94**	**0.174**	**0.0331**	**1.203**	**0.1065**

et al., 2020; Martin-Brualla et al., 2021; Pumarola et al., 2021). Therefore, we consider applying the positional encoding function[3] to our keypoints when we produce the scale-aware memory. We show the results in Tab .3. From the Tab. 3, we observe that when we apply the position encoding function on keypoints cannot bring the improvements, and even degrade the model if we set the L as 20. Since the keypoints will be utilized to estimate the motion flow in 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 bad generation.

The input element in ISCM. We also conduct experiments to investigate the usage of intermediate feature $\mathbf{F}^i_{proj}$ ("ISCM w/o $\mathbf{F}^i_{proj}$") and keypoints (ISCM w/o keypoints), the results shown in Tab. 4 indicate that these two items are equally crucial for the generation of the scale-aware memory bank. We can obtain the best results when we combine them together.

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

The dynamic convolution in MCM. Besides, we also conduct an ablation study on the dynamic convolution layer in the memory compensation module (see Fig.4). We can observe that the dynamic convolution layer can contribute to the final performance, especially for the AKD and AED.

Table 5: Comparisons results on VoxCeleb2.

Model	SSIM (%) ↑	PSNR ↑	LPIPS ↓	$\mathcal{L}_1$ ↓	AKD ↓	AED ↓
FOMM	77.19	30.71	0.257	0.0513	1.762	0.212
MRAA	78.07	30.89	0.262	0.0511	1.796	0.213
DaGAN	79.02	30.81	0.250	0.0483	1.865	0.341
TPSM	78.22	30.63	0.254	0.0527	1.703	0.210
Ours w/o ISCM	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**

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

Table 6: Comparisons results on HDTF dataset.

Model	SSIM (%) ↑	PSNR ↑	LPIPS ↓	$\mathcal{L}_1$ ↓	AKD ↓	AED ↓
FOMM	76.94	31.87	0.155	0.0363	1.116	0.092
MRAA	79.36	32.32	0.156	0.0331	1.039	0.100
DaGAN	82.29	32.29	0.136	0.0304	1.020	0.252
TPSM	86.05	32.85	0.114	0.0264	1.015	0.072
Ours w/o ISCM	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**

More datasets for evaluation. To fully verify the superiority of our method, we also compare it with other methods on two other large datasets, *i.e.* VoxCeleb2 (Chung et al., 2018) and HDTF (Zhang et al., 2021). We report the results on Tab. 5 and Tab. 6. From these two tables, we can observe that our method can still obtain the best results compared with other SOTA methods[4]. It strongly illustrates the superiority of our designed method.

Idenetity Preservation. In this section, we reorganized the voxceleb1 dataset and divided 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 Tab. 7. We can observe that our method obtains higher performance under the setting of testing identity as a part of the training corpus. One possible reason is that our global memory is learned from the identities in the training set. In this way, it can better compensate for the facial details of these seen identities.

Table 7: Comparisons results on HDTF dataset.

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**

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 demo videos, we can observe that our main can compensate those regions that do not appear in the source image better than other methods (*e.g.* the ear region in demo2). These demos are attached in Supplementary Material.

Comparison in other domains. To better verify the generalization ability of our method, we also train our method on TED-talks dataset (Siarohin et al., 2021), because the human body is also symmetrical and highly structured. We report the results in Tab. 8. From the Tab. 8, our method still obtain the best results among all compared method. This generalization experiments verity that our memory can learn the symmetrical and structured object to inpaint the generated image.

Table 8: Comparisons results on TED-talks dataset.

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

Meta memory visualization. In this section, we show all channels of our learn meta memory in Fig. 9 for better understanding. To look into details, we also show some channels in Fig. 10 in high resolution. These visualizations demonstrate the meaningful facial priors learnt in the meta memory.

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

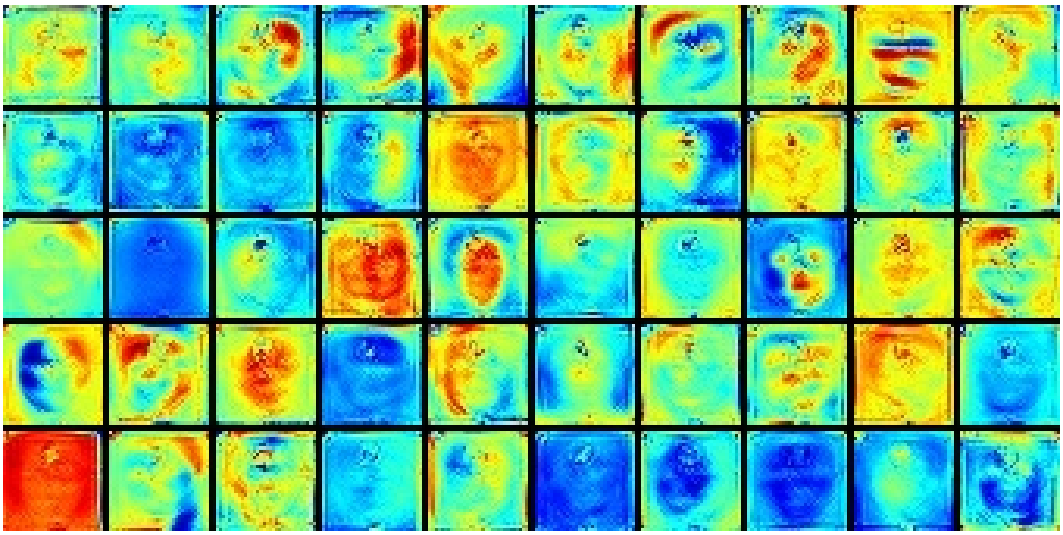


Figure 9: The visualization of randomly selected channels of the meta memory $\mathbf{M}_o$.

More qualitative ablation studies . To better show that our designed implicit scale conditioned memory module brings improvement to our model, we illustrate more qualitative results for ablation studies in Fig. 11 and Fig. 12.

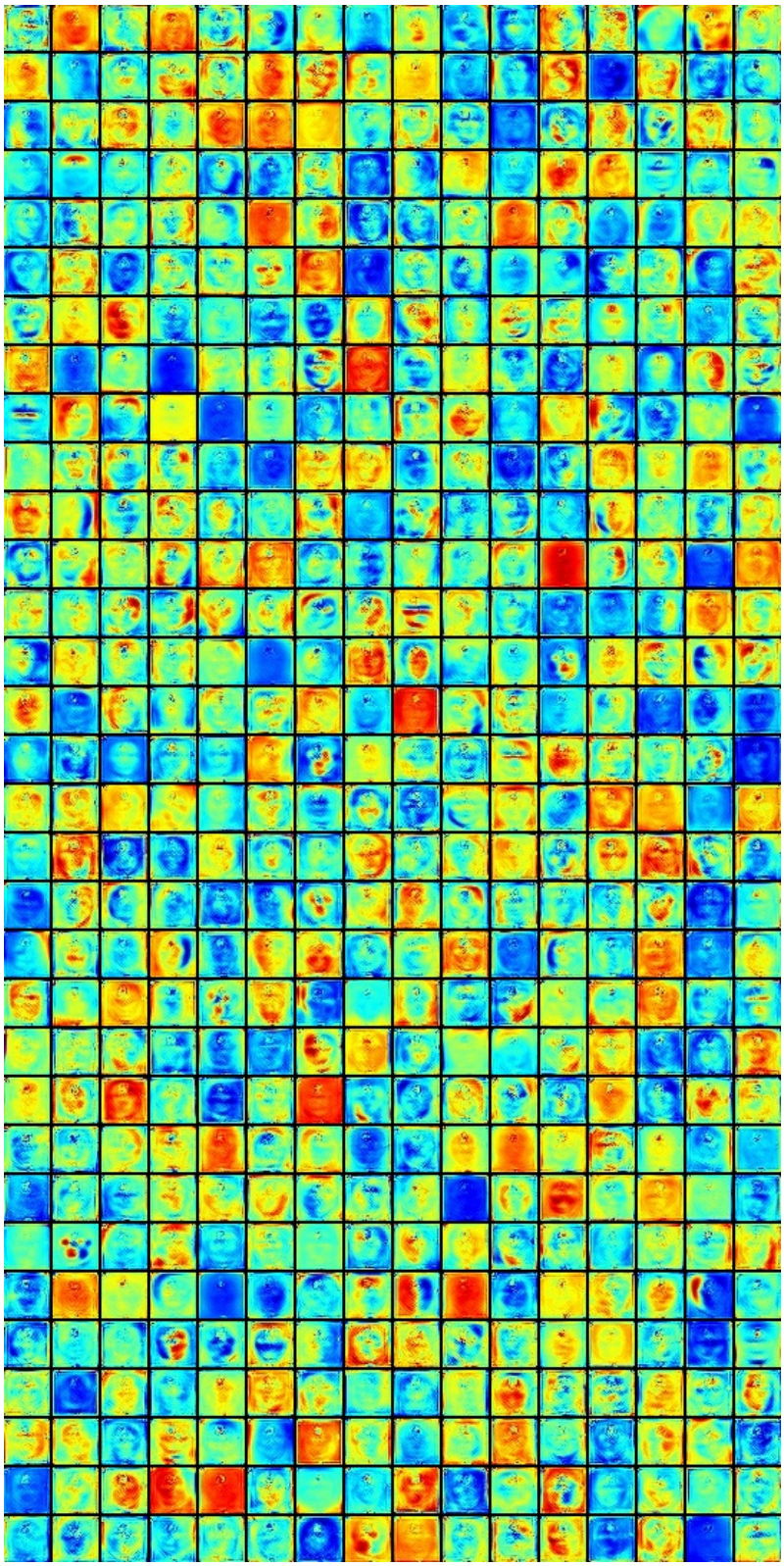


Figure 10: The visualization of all channels of the meta memory $\mathbf{M}_o$.

Figure 11: Qualitative ablation studies in VoxCeleb1 dataset. The memory compensation module (MCM) and implicit scale conditioned memory module (ISCM) can obtain improvements.

Figure 12: Qualitative ablation studies in VoxCeleb1 dataset. The memory compensation module (MCM) and implicit scale conditioned memory module (ISCM) can obtain improvements.