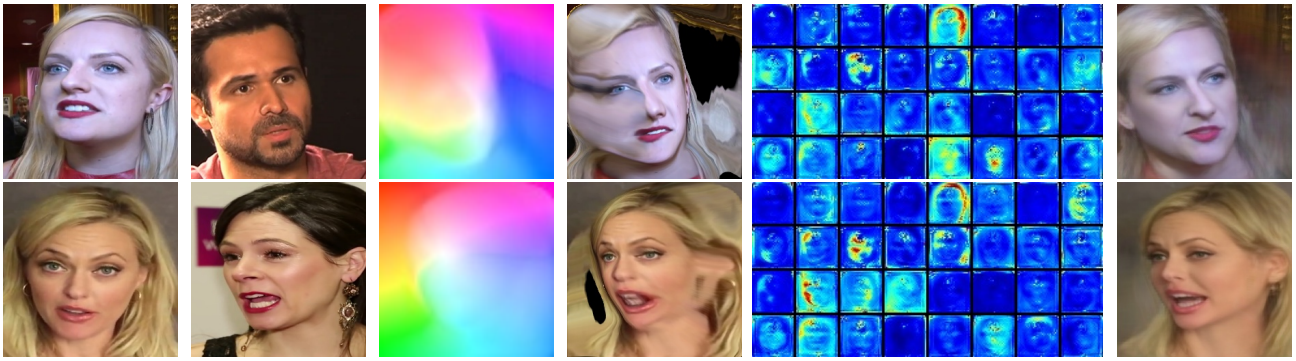
Implicit Identity Representation Conditioned Memory Compensation Network for Talking Head Video Generation

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

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



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

Figure 1: The animation illustration of the proposed implicit identity representation conditioned memory compensation network (MCNet). MCNet first learns the motion flow (**c**) between the source and driving images; (**d**) shows possible occlusion or deformation artifacts caused by large head motion. The warped images are produced by warping the source image with the motion flow; (**e**) presents randomly sampled memory channels of our learned memory bank conditioned with implicit-keypoint representations. Examples of generated results with our memory compensation network are shown in (**f**).

Abstract

Talking head video generation aims to animate a human face in a still image with dynamic poses and expressions using motion information derived from a target-driving video, while maintaining the person's identity in the source image. However, dramatic and complex motions in the driving video cause ambiguous generation, because the still source image cannot provide sufficient appearance information for occluded regions or delicate expression variations, which produces severe artifacts and significantly degrades the generation quality. To tackle this problem, we propose to learn a global facial representation space, and design a novel implicit identity representation conditioned memory compensation network, coined as MCNet, for high-fidelity talking head generation. Specifically, we devise a network module to learn a unified spatial facial meta-memory bank from all training samples, which can provide rich facial structure and appearance priors to compensate warped source facial features for the generation. Furthermore, we propose an effective query mechanism based on implicit identity representations learned from the discrete keypoints of the source image. It can greatly facilitate the retrieval of more correlated information from the memory bank for the compensation. Extensive experiments demonstrate that MCNet can learn representative and complementary facial memory, and can clearly outperform previous state-of-the-art talking head generation methods on VoxCeleb1 and CelebV datasets. Please check our Project.

1. Introduction

In this paper, we aim at addressing 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 presented 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 or 3D to improve the generation. More specifically, 2D facial keypoints or landmarks are learned to model the motion flow (see Fig. 1c) between the source image and any target image in the driving video [38, 36, 7]. Some works also consider utilizing 3D facial prior model (*e.g*. 3DMM[1]) with decoupled expression codes [38, 36] or learning dense facial geometries in a self-supervised manner [7] 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. 1d), because the still source image cannot provide sufficient appearance information for occluded regions or delicate expression variations, which severely produces artifacts and significantly degrades the generation quality.

Intuitively, we understand that human faces are highly symmetrical and 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 [16], 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 is highly potential for high-fidelity talking head generation, while it has been barely explored in existing works.

In this paper, to effectively deal with the ambiguities in dramatic appearance changes from the still source image, we propose an implicit identity representation conditioned **M**emory **C**ompensation **Net**work, coined as **MCNet**, to learn and transfer global facial representations to compensate ambiguous facial details for a high-fidelity generation. Specifically, we design and learn a global and spatial facial 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 most representative facial patterns globally. Since the different source face images contain distinct structures and appearances, to more effectively query the learned global meta memory bank, we propose an implicit identity representation conditioned memory module (IICM) (see Fig. 3). The implicit identity representation is learned from both the discrete keypoint coordinates of the source face image that contains the facial structure information, and the warped source feature map that represents facial appearance distribution. Then, we further use it to condition the query on the global facial meta-memory bank to learn a more correlated memory bank for the source, which can effectively compensate the source facial feature maps for the generation. The compensation is then performed through a proposed memory compensation module (MCM) (see Fig. 4).

We conduct extensive experiments to evaluate the proposed MCNet on two competitive talking head generation datasets (*i.e*. VoxCeleb [15] and CelebV [29]. 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 over state-of-the-art methods from both qualitative and quantitative perspectives.

In summary, our main contribution is three-fold:

- We propose to learn a global facial meta-memory bank to transfer representative facial patterns to handle the appearance and structure ambiguities caused by 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 to address the ambiguities in talking head generation.

- We propose a novel implicit identity representation conditioned memory compensation network (MCNet) for talking head video generation, in which an implicit identity representation conditioned memory module (IICM) and a facial memory compensation module (MCM) are designed to respectively perform the meta-memory query and feature compensation.

- 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 approach can effectively boost the performance of different talking head generation frameworks.

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 im-

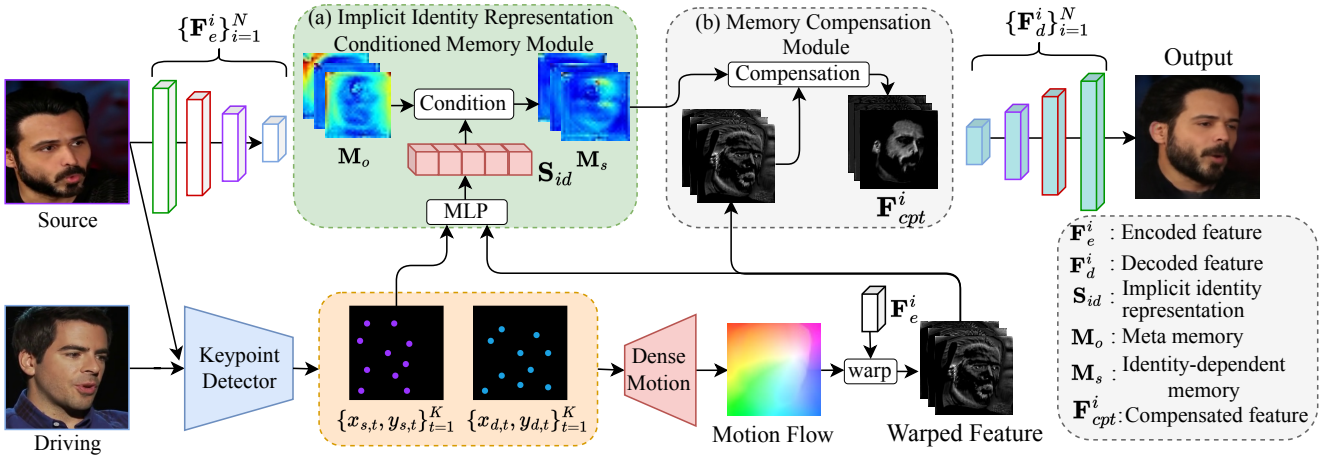


Figure 2: An overview of the proposed MCNet. It contains two designed modules to compensate the source facial feature map: (i) The implicit identity representation conditioned memory module (**IICM**) learns a global facial meta-memory bank, and an implicit identity representation from facial keypoint coordinates of the source image, which conditions on the query of the learned meta-memory bank, to obtain more structure-correlated facial memory to the warped source feature map for compensation; (ii) The memory compensation module (**MCM**) designs a dynamic cross-attention mechanism to perform a spatial compensation for the warped source feature map for the generation.

age. Several approaches [33, 30, 23] utilized a 3DMM regressor [21, 39] 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, some other works [22, 6, 35, 36, 38] utilized facial landmarks detected by a pretrained face model [5] to act as anchors of the face. Then, the facial motion flow calculated from the 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 [17, 7, 24, 13, 37] 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 [9, 14, 28, 10] 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*. [12] explicitly divide the driving audio into granular parts through delicate priors to control the lip shape, face pose, and 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, which can effectively compensate intermediate facial features and produce higher-quality generation on ambiguous regions caused by large head motion.

Memory Bank Learning. Introducing an external memory or prior component is popular because of its flexible capability of storing, abstracting, and organizing long-term knowledge into a representative form. Recently, the memory bank has shown its powerful capabilities in learning and reasoning for addressing several challenging tasks, *e.g*. image processing [34, 8], video object detection [19], and image caption [4]. As an earlier work, [26] proposes 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. Xu *et al*. [32] use the texture memory of patch samples extracted from unmasked regions to inpaint missing facial parts. [31] 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 learning a global facial meta-memory bank to deal with ambiguous generation issues in the task of talking head generation. We also accordingly design a novel implicit identity representation conditioned memory query mechanism and a memory compensation network to effectively tackle the issues.

3. The Proposed Approach

3.1. Overview

An overview of our proposed implicit identity representation conditioned memory compensation network for talking head generation is depicted in Fig. 2. It 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

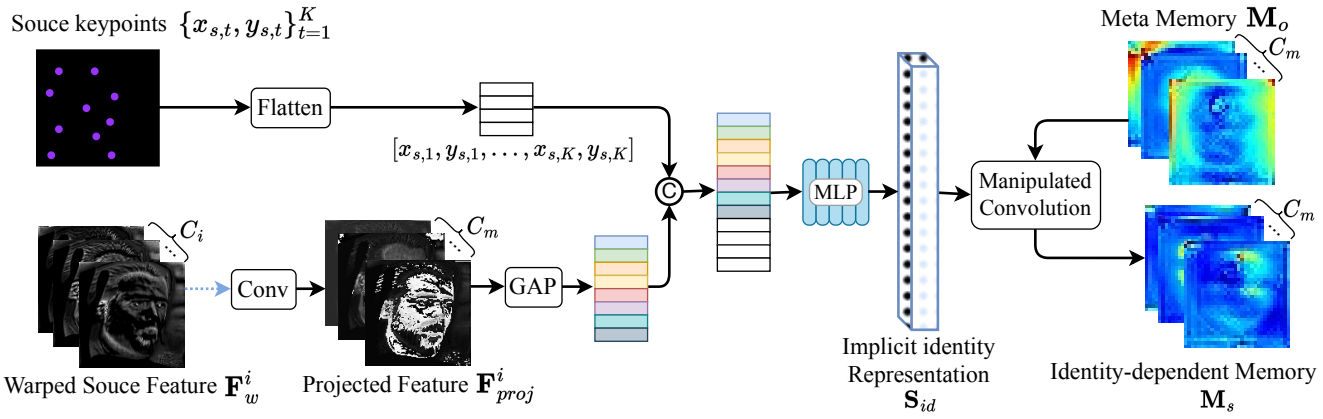


Figure 3: Illustration of the proposed implicit identity representation conditioned memory module (IICM). The symbol © denotes the concatenation operation, and the "GAP" and "Conv" represent the global average pooling and the convolution layer, respectively. The detailed generation of the projected feature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

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

3.2. Learning Implicit Identity Representation Conditioned Global Facial Meta-memory

We first aim at learning a global meta-memory bank to model facial structure and appearance representations from the whole face dataset. As different human faces have distinct structures and appearances, and thus using the whole learned facial meta-memory bank to directly compensate different source faces is inflexible. To handle this issue, we further learn an implicit identity representation from discrete keypoint coordinates of the source face and the corresponding warped source feature map. It is then used to condition on the query of the global meta-memory bank and obtain source-identity-dependent feature memory, which compensates the warped source feature map for generation.

Global facial meta-memory. In this work, we first aim to learn a global facial meta-memory bank to store the global and representative facial appearance and structure priors 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 [3]. Moreover, the multiple channels hold sufficient capacity for the meta-memory bank 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. The meta memory bank is automatically updated by the optimization gradients from all the training images during the training stage, based on an objective function described in Eq. 4. In this way, the facial prior learned in the meta memory is global rather than conditioned on any specific input sample, providing highly beneficial global patterns for compensating face generation.

Implicit identity representation learning. In our framework, the detected facial keypoints are used to learn motion flow for feature warping. The facial keypoints implicitly contain the structure information of the face because of their structural positions [17, 20]. Therefore, we utilize both the source keypoint coordinates $\{x_{s,t}, y_{s,t}\}_{t=1}^{K}$ and its corresponding warped feature $\mathbf{F}_w^i$ that provides additional appearance constraints to learn an implicit identity representation of the source face. The reason for learning on the source is that we need to compensate the warped facial feature map with the identity of the source. As shown in Fig. 3, we first utilize a global average pooling function $\mathcal{P}_F$

to squeeze the global spatial information of the projected feature $\mathbf{F}^i_{proj}$ that is produced from the warped feature $\mathbf{F}^i_w$ (see Fig. 4), into a channel descriptor. It is then concatenated with flattened and normalized keypoint coordinates, and fed into an MLP mapping network $\mathcal{F}_{mlp}$ to learn an implicit identity representation of the source image. We have:

$$\mathbf{S}_{id} = \mathcal{F}_{mlp}\left(\left[\mathcal{P}_F(\mathbf{F}^i_{proj}), [x_{s,1}, y_{s,1}, \ldots, x_{s,K}, y_{s,K}]\right]\right),$$

where the operator $[\cdot,\cdot]$ indicates the concatenation operation, and $\mathbf{S}_{id}$ is the learned implicit identity representation of the source face image.

implicit identity representation conditioned meta-memory learning. As discussed before, human faces present distinct structures and appearances. To generate a more correlated facial memory for compensating the source feature map, we utilize the learned implicit identity representation $\mathbf{S}_{id}$ to condition on the retrieval of our global facial meta-memory $\mathbf{M}_o$, which produces an identity-dependent facial memory $\mathbf{M}_s$ for each source face image. Inspired by the style injection in StyleGANv2 [11], we utilize the implicit identity representation $\mathbf{S}_{id}$ to manipulate a 3×3 convolution layer to produce a conditioned 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 identity representation $\mathbf{S}_{id}$, and j and k enumerate the output feature maps and spatial footprint of the convolution, respectively. Finally, we obtain the learned source-dependent facial memory:

$$\mathbf{M}_s = \mathcal{F}_{C_{\omega''}}(\mathbf{M}_o) \tag{1}$$

where the $\mathcal{F}_{C_{\omega''}}$ is the manipulated convolution layer parameterized by ω''. With the identity-independent memory $\mathbf{M}_s$, each warped source feature map can be compensated by the source-correlated facial priors to have a more effective face generation.

3.3. Global Memory Compensation and Generation

The warped source feature map typically contains ambiguity for the generation, especially when the warping is performed under large head motion or occlusion. Thus, we propose to inpaint those ambiguous features via compensating the warped source facial feature maps. To this end, we design a memory compensation module (MCM) as shown in Fig. 4, to refine the warped feature $\mathbf{F}^i_w$ via the learned source-identity-dependent facial memory bank $\mathbf{M}_s$.

Projection of warped facial feature. To maintain better the identity information in the source image while compensating the warped source feature map, we employ a channel-split strategy to split the warped feature $\mathbf{F}^i_w$ into two parts along the channel dimension, *i.e.* $\mathbf{F}^{i,0}_w$ and $\mathbf{F}^{i,1}_w$. The first half of channels $\mathbf{F}^{i,0}_w$ are directly passed through for contributing the identity preserving, while the rest half of channels $\mathbf{F}^{i,1}_w$ are modulated by the source identity-dependent memory bank $\mathbf{M}_s$, to refine the ambiguities. After splitting, we employ a 1×1 convolution layer on $\mathbf{F}^{i,1}_w$ to change the channel number, resulting in a projected feature map $\mathbf{F}^i_{proj}$.

Warped facial feature compensation. We adopt a dynamic cross-attention mechanism to compensate the warped source feature map spatially. Specifically, we employ the identity-dependent memory $\mathbf{M}_s$ to produce the Key $\mathbf{F}^i_K$ and Value $\mathbf{F}^i_V$ via two dynamic convolution layers (*i.e.* f^1_{dc}, f^2_{dc}) conditioned on the projected feature $\mathbf{F}^i_{proj}$. In this way, the generated Key and Value are identity-dependent and capable of providing useful context information. In the meanwhile, we perform a non-linear projection to map $\mathbf{F}^i_{proj}$ into a query feature $\mathbf{F}^i_Q$ by a 1×1 convolution layer followed by a ReLU layer. Then, we perform cross attention to reconstruct a more robust feature $\mathbf{F}^i_{ca}$ as:

$$\mathbf{F}^i_{ca} = \mathcal{F}_{C_{1\times1}}\left(\text{Softmax}\left({\mathbf{F}^i_Q}^T \times \mathbf{F}^i_K\right) \times \mathbf{F}^i_V\right), \tag{2}$$

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 $\mathbf{F}^i_{ca}$ with the first-half channels $\mathbf{F}^{i,0}_w$:

$$\mathbf{F}^i_{cpt} = \text{Concat}[\mathbf{F}^i_{ca}, \mathbf{F}^{i,0}_w], \tag{3}$$

where the $\text{Concat}[\cdot,\cdot]$ represents a concatenation operation. As a result, the final output feature map $\mathbf{F}^i_{cpt}$ can effectively benefit and incorporate the learned facial prior information [25] from the memory, modulated by the dynamic cross-attention mechanism.

Regularization on consistency. To learn the global facial appearance and structure representations from the input training face images, we need to make the learning of the meta-memory constrained by every single image in the training data. Simply but effectively, we enforce the consistency between the projected feature $\mathbf{F}^i_{proj}$ from the current training face image, and the value feature $\mathbf{F}^i_V$ from the global meta memory:

$$\mathcal{L}_{con} = ||\mathbf{F}^i_V - de(\mathbf{F}^i_{proj})||_1, \tag{4}$$

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 into the learning of the global meta-memory, while not affecting the learning of the source image features. This can guarantee the stability of training the overall generation framework. The above equation also makes sure that the optimization gradients from all the face

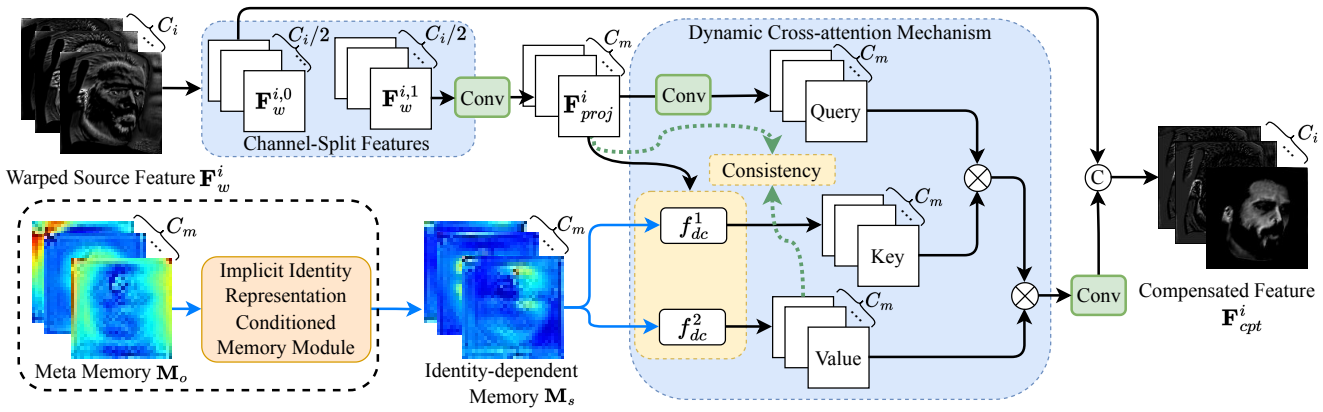


Figure 4: The illustration of the memory compensation module (MCM). The symbol $\otimes$ denotes matrix multiplication, and f_{dc}^1 and f_{dc}^2 are dynamic convolution layers [2], whose kernel weights are estimated by the projected feature $\mathbf{F}_{proj}^i$. The © represents the concatenation operation, 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.

images during the training state contribute together to the updating of the memory bank, and thus it can capture global facial representations for the generation compensation.

Multi-layer generation. Following previous works [37] considering multi-scale features in generation, we also perform memory compensation for feature maps of multiple layers to enhance facial details. 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 IICM and MCM modules sequentially to produce the compensated feature maps $\{\mathbf{F}_{cpt}^i\}_{i=1}^N$. 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 the i-th level ($i > 1$), the output compensated feature map $\mathbf{F}_{cpt}^i$ is concatenated with the decoded feature $\mathbf{F}_d^i$ at the same level to produce a decoded feature $\mathbf{F}_d^{i+1}$. 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 $\mathbf{M}_o$.

3.4. Training

We train the proposed MCNet by minimizing several optimization losses. Similar to FOMM [17], 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 $\mathcal{L}_{dist}$ [7] to avoid the detected keypoints crowding around a small neighborhood. The $\mathcal{L}_{con}$ is the regularization consistency loss described in Eq. 4. The overall loss function is written as follows:

$$\mathcal{L} = \lambda_P \mathcal{L}_P + \lambda_{eq}\mathcal{L}_{eq} + \lambda_{dist}\mathcal{L}_{dist} + \lambda_{con}\mathcal{L}_{con}, \quad (5)$$

where the λ_P, λ_{eq}, λ_{dist} and λ_{con} are the hyper-parameters to allow for balanced learning from these losses. More details about these losses are described in Supplementary.

4. Experiments

In this section, we present quantitative and qualitative experiments to validate the effectiveness of our MCNet.

4.1. Datasets and Metrics

Dataset. We evaluate our MCNet on two talking head generation datasets, *i.e.* VoxCeleb1 [15] and CelebV [29] dataset. We follow the sampling strategy for the test set in DaGAN [7] 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 [17], 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 [7], 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 [17] and DaGAN [7], which adopt the same

(a) Same-identity Reenactment (b) Cross-identity Reenactment

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.

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 [27]	71.9	22.54	-	0.0780	7.687	0.405	-	-	0.679	3.62
marioNETte [6]	75.5	23.24	-	-	-	-	-	-	0.710	3.41
FOMM [17])	72.3	30.39	0.199	0.0430	1.294	0.140	0.882	2.824	0.667	3.90
MeshG [33]	73.9	30.39	-	-	-	-	-	-	0.709	3.41
face-vid2vid [24]	76.1	30.69	0.212	0.0430	1.620	0.153	0.839	4.398	0.805	3.15
MRAA [18]	80.0	31.39	0.195	0.0375	1.296	0.125	0.882	2.751	0.840	2.46
DaGAN [7]	80.4	31.22	0.185	0.0360	1.279	0.117	0.888	2.822	0.873	2.33
TPSN [37]	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**

Table 1: Comparisons with state-of-the-art methods on same-identity reenactment on VoxCeleb1 (see Fig. 5a) and cross-identity reenactment on VoxCeleb1 and CelebV dataset (see Fig. 5b).

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. 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 coun-

Figure 6: Qualitative ablation studies. The memory compensation module (MCM) and implicit identity representation conditioned memory module (IICM) can both effectively improve the generation performances. The last column verifies that our IICM can learn identity-conditioned memories (*i.e.* $\mathbf{M}_s$) for the different source face samples.

Model	SSIM (%) ↑	PSNR ↑	LPIPS ↓	$\mathcal{L}_1$ ↓	AKD ↓	AED ↓
Baseline	81.1	31.70	0.182	0.0356	1.303	0.124
Baseline + MCM$^{\text{w/o Eq. 3}}$	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 + IICM + MCM (MCNet)	**82.5**	**31.94**	**0.174**	**0.0331**	**1.203**	**0.106**
FOMM [17]	72.3	30.39	0.199	0.0430	1.294	0.140
FOMM+ IICM + MCM	**81.8**	**31.73**	**0.179**	**0.0353**	**1.269**	**0.119**
TPSN [37]	81.6	31.43	0.179	0.0365	1.233	0.119
TPSN+ IICM + MCM	**82.0**	**31.55**	**0.175**	**0.0356**	**1.216**	**0.115**

Table 2: Ablation studies: "Baseline" indicates the simplest model without the implicit identity representation conditioned memory module (IICM) and memory compensation module (MCM). "MCM$^{\text{w/o Eq. 3}}$" indicates that we use the entire warped feature to generate the projected feature $\mathbf{F}^i_{proj}$ without using the channel split. The compensated feature $\mathbf{F}^i_{cpt}$ for generation is thus directly from the output of cross-attention query (*i.e.* $\mathbf{F}^i_{ca}$) of the meta memory $\mathbf{M}_o$.

tenance 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 identity representation conditioned memory module (IICM) and memory compensation module (MCM). We report their quantitative results in Table 2 and the qualitative results in Fig. 6. Our baseline is the model without IICM and MCM modules. The "Baseline + MCM" means that we drop the IICM module and replace the identity-independent memory $\mathbf{M}_s$ with the meta memory $\mathbf{M}_o$ shown in Fig. 4.

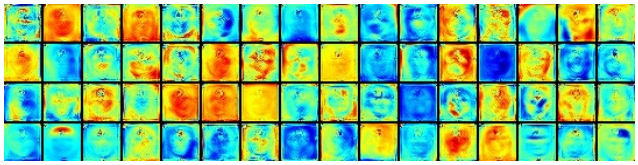


Figure 7: The visualization of randomly selected channels of the global meta memory $\mathbf{M}_o$. It can be observed that our meta-memory learns very diverse facial representations.

Effect of meta memory learning. We first visualize the learned meta memory in Fig. 7, which aims to learn the globally representative facial appearance and structure pat-

terns. In Fig. 7, we visualize partial channels of the facial meta-memory bank. It can be observed that these channels represent faces with distinct appearances, structures, scales, and poses, which are very informative and clearly beneficial for facial compensation and generation, confirming our motivation of learning global facial representations to tackle ambiguities in the talking head generation.

Effect of 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 a comparison between "Baseline+MCM" and "Baseline" (82.3% vs. 81.1% on SSIM). In Fig. 6, we can also 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 + $\mathrm{MCM}^{\mathrm{w/o\,Eq.3}}$" indicate 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 map to produce higher-fidelity results for the talking head generation.

Effect of implicit identity representation conditioned memory learning. To verify the effectiveness of the proposed implicit identity representation conditioned memory module (*i.e.* IICM introduced in Sec. 3.2), we show the randomly sampled channels of the conditioned memory bank in Fig 6. As shown in the last column in Fig 6, the IICM produces an identity-dependent memory bank for the input source images. By deploying the IICM module, our MCNet can generate highly realistic-looking images compared with "Baseline+MCM", verifying that the learned memory conditioned on the input source provides a more effective compensation on the warped source feature map for the talking face generation.

The effect of consistency regularization in Eq. 4. Eq. 4 we introduced ensures that the optimization gradients across all facial images collaborate in updating the memory bank, capturing global facial representations to improve face generation. Without it, as shown in Fig. 8, the facial memory bank clearly learns less discriminative face patterns and more noisy representations.

Generalization experiment. Importantly, we also embed the proposed MCM and IICM modules into different representative talking head generation frameworks, including FOMM [17] and TPSM [37], to verify our designed memory mechanism can be flexibly generalized to existing models. As shown in Table 2, the TPSM, which has a different motion estimation method compared to ours, with a deployment of our proposed memory modules, can achieve a stable improvement. The "FOMM+IICM+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. Additionally, we show our generation results on some out-of-domain samples in Fig. 9, to verify the out-of-domain generation capabilities of our model. As shown in Fig. 9, our method is able to effectively modify the expression of the oil-painted and cartoon faces.

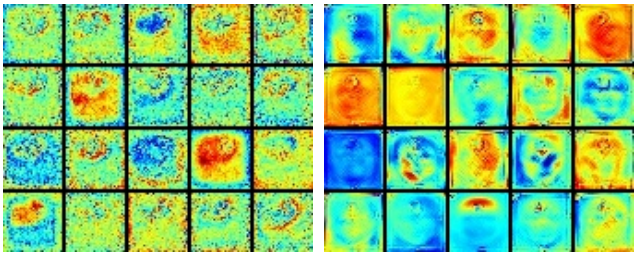


Figure 8: Visualization of selected channels of the meta memory from our full method and the ablation method (*i.e.*, w/o consistency regularization introduced in Eq. 4).

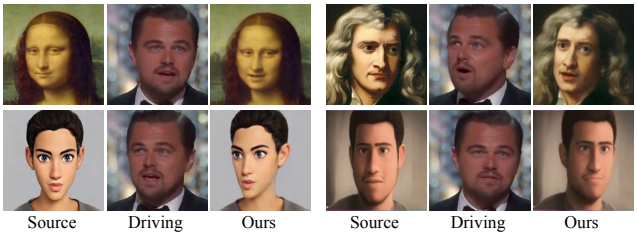


Figure 9: Qualitative results of out-of-domain generation.

5. Conclusion

In this paper, we present an implicit identity representation conditioned memory compensation network (MCNet) to globally learn representative facial patterns to address the ambiguity problem caused by the dynamic motion in the talking head video generation task. MCNet utilizes a designed implicit identity representation conditioned memory module to learn the identity-dependent facial memory, which is further used to compensate the warped source feature map by a proposed memory compensation module. Extensive results clearly show the effectiveness of learning global facial meta-memory for the task, producing higher-fidelity results compared with the state-of-the-arts.

6. Acknowledgement

This research is supported in part by HKUST-SAIL joint research funding, the Early Career Scheme of the Research Grants Council (RGC) of the Hong Kong SAR under grant No. 26202321, and HKUST Startup Fund No. R9253.

References

[1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999. 2

[2] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *CVPR*, 2020. 6

[3] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 4

[4] Zhengcong Fei. Memory-augmented image captioning. In *AAAI*, 2021. 3

[5] 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. 3

[6] Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *AAAI*, 2020. 3, 7

[7] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *CVPR*, 2022. 2, 3, 6, 7

[8] Huaibo Huang, Aijing Yu, and Ran He. Memory oriented transfer learning for semi-supervised image deraining. In *CVPR*, 2021. 3

[9] 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. In *SIGGRAPH*, 2022. 3

[10] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *CVPR*, 2021. 3

[11] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 5

[12] 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. 3

[13] 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*, 2021. 3

[14] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: real-time photorealistic talking-head animation. *TOG*, 2021. 3

[15] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. In *INTERSPEECH*, 2017. 2, 6

[16] Haibo Qiu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. End2end occluded face recognition by masking corrupted features. *TPAMI*, 44(10):6939–6952, 2021. 2

[17] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *NeurIPS*, 2019. 3, 4, 6, 7, 8, 9

[18] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021. 7

[19] Guanxiong Sun, Yang Hua, Guosheng Hu, and Neil Robertson. Mamba: Multi-level aggregation via memory bank for video object detection. In *AAAI*, 2021. 3

[20] 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. 4

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

[22] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Facegan: Facial attribute controllable reenactment gan. In *WACV*, 2021. 3

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

[24] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021. 3, 7

[25] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 5

[26] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014. 3

[27] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *ECCV*, 2018. 7

[28] 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 MM*, 2021. 3

[29] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *ECCV*, 2018. 2, 6

[30] 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*, 30:8658–8670, 2021. 3

[31] Zhuojie Wu, Xingqun Qi, Zijian Wang, Wanting Zhou, Kun Yuan, Muyi Sun, and Zhenan Sun. Showface: Coordinated face inpainting with memory-disentangled refinement networks. In *BMVC*, 2022. 3

[32] Rui Xu, Minghao Guo, Jiaqi Wang, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Texture memory-augmented deep patch-based image inpainting. *TIP*, 30:9112–9124, 2021. 3

[33] Guangming Yao, Yi Yuan, Tianjia Shao, and Kun Zhou. Mesh guided one-shot face reenactment using graph convolutional networks. In *ACM MM*, 2020. 3, 7

[34] 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. 3

[35] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *ECCV*, 2020. 3

[36] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, 2019. 2, 3

[37] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *CVPR*, 2022. 3, 6, 7, 8, 9

[38] Ruiqi Zhao, Tianyi Wu, and Guodong Guo. Sparse to dense motion transfer for face image animation. In *ICCV*, 2021. 2, 3

[39] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *TPAMI*, 41(1):78–92, 2017. 3