Learning Flow-Based Disentanglement

Jen-Tzung Chien, *Senior Member, IEEE*, and Sheng-Jhe Huang

***Abstract*— Face reenactment aims to generate the talking face images of a target person given by a face image of source person. It is crucial to learn latent disentanglement to tackle such a challenging task through domain mapping between source and target images. The attributes or talking features due to domains or conditions become adjustable to generate target images from source images. This article presents an information-theoretic attribute factorization (AF) where the mixed features are disentangled for flow-based face reenactment. The latent variables with flow model are factorized into the attribute-relevant and attribute-irrelevant components without the need of the paired face images. In particular, the domain knowledge is learned to provide the condition to identify the talking attributes from real face images. The AF is guided in accordance with multiple losses for source structure, target structure, random-pair reconstruction, and sequential classification. The random-pair reconstruction loss is calculated by means of exchanging the attribute-relevant components within a sequence of face images. In addition, a new mutual information flow is constructed for disentanglement toward domain mapping, condition irrelevance, and condition relevance. The disentangled features are learned and controlled to generate image sequence with meaningful interpretation. Experiments on mouth reenactment illustrate the merit of individual and hybrid models for conditional generation and mapping based on the informative AF.**

***Index Terms*— Disentangled features, domain mapping, face reenactment, flow model, information-theoretic generation.**

I. Introduction

DOMAIN mapping aims to characterize the complicated relation between source and target domains where the conditional generation of target data can be learned with the specific embeddings of features, styles, or attributes from source data. It is essential to learn such a generative model that acts as the observation probability for generation of new samples. Basically, the deep generative models, combining generative models with deep neural networks, have been recognized as a building block in implementation of various multimedia information systems. Deep generative models in domain mapping have been developed for different pairs of mapping in the presence of various data types. The mapping pairs can also be under the same data type. For example, image-to-image translation is recognized as a popular domain mapping task for style transfer where the style of source images is transferred and incorporated in target images [1]–[3]. In general, a key success to generative model for domain mapping relies on the preservation of overall data structure based on the disentanglement in latent representation. Latent disentanglement typically aims to strengthen the learning representation by disentangling the basic structure of observations into disjoint components or salient features in the latent variable model. Currently, there is no clear definition and solution to latent disentanglement because the ground truth of disentangled features from the structural and mixed observations is missing. Nevertheless, the disentanglement needs to preserve the properties of independence as well as interpretation in latent representation [4]. Independence is to identify the statistically independent factors that are not interfered with each other, while the interpretation is to capture the semantic meanings of the separated components. The more the generative model understands the observations, the better the rich samples are precisely generated.

A. Related Work

Generative models are developed as the probabilistic distributions to reproduce or even create new data like a human does. Traditionally, latent disentanglement was developed for generative models based on the variational autoencoder (VAE) [5], [6], which depended on the expressiveness of prior distribution, and the generative adversarial network (GAN) [7], [8], which required the stability in minimax optimization for adversarial training. VAE also suffered from the posterior collapse [9] in variational inference, which resulted in the blurred data. As a result, the goodness of disentanglement affected the controllable generation for domain mapping. In [10] and [11], the richness of generated images was attained by the controllable generation where the styles and attributes were identified via transfer learning. Generation of face images is viewed as a popular task for image-to-image translation. The attributes of genders, races, hairstyles, and facial emotions were disentangled to pursue the variety of the generated faces. In [12], the latent semantics of facial expressions were adjusted to manipulate the attribute mapping in local regions where the face landmark was constructed for image synthesis [13]. In [14] and [15], talking information of a target video was combined with a source image to implement the talking face generation where GAN was applied for conditional generation with an audio-visual disentangled representation.

A recent paradigm, called flow-based model [16], [17], has achieved the state-of-the-art performance for generation of various types of data, including face images [18], medical images [19], natural sentences [20], [21], and speech waves [22]. The attractiveness of flow model is the exact

Manuscript received 18 December 2021; revised 16 April 2022 and 18 June 2022; accepted 7 July 2022. Date of publication 15 July 2022; date of current version 6 February 2024. This work was supported in part by the Ministry of Science and Technology, Taiwan, under Contract MOST 110-2634-F-A49-003. *(Corresponding author: Jen-Tzung Chien.)*

The authors are with the Department of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan (e-mail: jtchien@nycu.edu.tw).

Digital Object Identifier 10.1109/TNNLS.2022.3190068

© 2022 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

estimation of target distribution for highly nonlinear domain mapping through multiple simple and invertible transformations. However, the conditional generation using a flow-based model was hard to implement because there was no reconstruction imposed in the generation phase and only the inference phase adopted the model. In [18], the post-preprocessing mechanism was proposed as an alternative to indirectly impose the labels for semantic adjustment. In [23], the effective architecture for generative flow using the masked convolution was proposed. In [19], the dual invertible networks were exploited and learned for flow-based modality transfer. In [24] and [25], the guided images were generated by the conditional flow-based models for image colorization and edge detection.

B. Main Idea of This Work

This article presents the flow-based disentanglement for conditional generation in face reenactment. Such a flow model is suitable for precise reconstruction due to the invertibility in mapping between the observed domain and the latent domain. There are two approaches. The first approach is based on the attribute factorization (AF) flow where the disentanglement is preserved by exploring the structural features in consecutive changes given by an image sequence. The attribute-relevant and attribute-irrelevant encoders are introduced to identify facial features in the flow-based model, which represents a specific talking attribute and an overall latent structure of talking mouth, respectively. These two encoders are mutually collaborated and estimated according to the objectives for disentangled domain mapping in mouth reenactment. The second approach is called the mutual information (MI) flow, which conducts information-theoretic learning for flow transformation that consolidates the disentanglement to connect the relations between image sequences and latent variables. MI is optimized to build the flow-based model so as to disentangle the informative features. The attributes in facial features are retrieved and treated as the conditions for face generation. MI flow is implemented by using an invertible 1×1 convolution (known as the Glow) [18] where high-quality image synthesis is assured. A conditional prior distribution is additionally learned to express implicit talking attribute from face data. This article further handles the dimensional waste in latent vectors and preserves the capability of data compression in construction of face generative model. The physical meaning of latent disentanglement becomes intuitive and interpretable. The proposed AF flow and MI flow can be merged to reinforce the performance. In the experiments on face reenactment, the synthesis of talking mouth from different domains of images and the image reconstruction from the disentangled features are illustrated. The remaining of this article is organized as follows. Section II addresses the flow-based representation and disentanglement. Sections III and IV present the conditional generation based on the disentangled flow models using AF flow and MI flow, respectively. Section V reports a series of experiments to evaluate these methods. The final conclusions drawn in this study are given in Section VI.

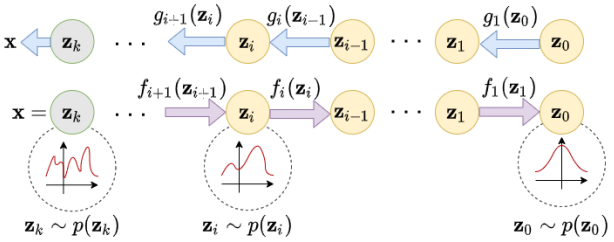

Fig. 1. Generating and normalizing processes in flow model.

II. Background Survey

First, the flow-based generative models for image-to-image translation with latent disentanglement are introduced.

A. Flow-Based Representation

Flow-based representation [16] was proposed as a new type of likelihood-based generative model where the distribution due to invertible transformation $\mathbf{z} = f_\theta(\mathbf{x})$ with parameter θ from observed sample $\mathbf{x}$ to latent variable $\mathbf{z}$ is calculated by $p_z(\mathbf{z}) = p_x(f_\theta^{-1}(\mathbf{z}))|\det((df_\theta^{-1}(\mathbf{z}))/d\mathbf{z})|$, where f_θ has an inverse function g_θ, i.e., $g_\theta(\mathbf{z}) = f_\theta^{-1}(\mathbf{z}) = \mathbf{x}$, and $\mathbf{J}_{f_\theta^{-1}} = ((df_\theta^{-1}(\mathbf{z}))/d\mathbf{z})$ denote the Jacobian matrix of f_θ^{-1} or g_θ. This normalizing process is reverse to form the generative process using $p_x(\mathbf{x}) = p_z(g_\theta^{-1}(\mathbf{x}))|\det((dg_\theta^{-1}(\mathbf{x}))/d\mathbf{x})|$. Fig. 1 shows the generative (upper) process and the normalizing (lower) process in flow-based model where a series of invertible functions $\{g_i(\mathbf{z}_{i-1})\}_{i=1}^k$ and $\{f_i(\mathbf{z}_i)\}_{i=1}^k$ is estimated to smoothly generate from $\mathbf{z}_0$ to $\mathbf{x} = \mathbf{z}_k$ and normalize from $\mathbf{x} = \mathbf{z}_k$ to $\mathbf{z}_0$, respectively, where the dimension of different variables $\{\mathbf{z}_i\}_{i=0}^k$ is fixed. The observed variable $\mathbf{x} = \mathbf{z}_k$ and latent variable $\mathbf{z}_0$ are represented by the complex distribution and the simple distribution (i.e., standard Gaussian), respectively. The flow-based generative model turns out to estimate the flow parameter θ by minimizing the expected loss function given by an exact likelihood-based model

$$\mathcal{L}_f(\mathbf{x};\theta) = -\mathbb{E}_{\mathbf{x}}\left[\log p_\theta(\mathbf{z}) + \log\left|\det\frac{df_\theta(\mathbf{x})}{d\mathbf{x}}\right|\right]. \quad (1)$$

After the first stage of flow-based pretraining, the second stage is devoted to find a feature encoder for AF. In [16], nonlinear independent component estimation (NICE) provided the transformation f_θ, which was easy to compute the inverse $f_\theta^{-1}(\mathbf{z})$ and the Jacobian determinant $\det((df_\theta(\mathbf{x}))/d\mathbf{x})$. The volume-preserving flow was built by imposing the additive coupling layer where the unit Jacobian determinant was obtained to assure volume preserving. However, such a flow is difficult to handle high-dimensional continuous space. Therefore, the real-valued nonvolume-preserving (RealNVP) flow [26] was proposed by implementing the affine coupling layer with masked convolution and multiscale architecture. In [18], Glow was inherited from an RealNVP multiscale structure that was driven by invertible 1×1 convolution. Flow models have been successfully developed for computer vision [19], [27] and natural language processing [20], [22].

B. Conditional Generation and Mapping

This article presents the flow-based domain mapping via conditional likelihood $p(\mathbf{y}|\mathbf{x})$ with a latent representation

$\mathbf{z} = f_\theta(\mathbf{x})$, which generates data $\mathbf{y}$ in the target domain conditioned on the source samples $\mathbf{x}$. Talking face reenactment conducts a kind of domain mapping, which generates the face images with lip motion via video frames [28]. This task aims to generate the talking face images of a target person given by a reference face of a source person. The challenges of this task are caused by the richness of lip movements and the sparseness of paired samples. Traditional method to handle this work was based on the 3-D face structural model [29], the dense photometric consistency measure [30], or the facial embedding representation [31]. More recently, the GAN was employed in domain separation and adaptation where adversarial learning was adopted to improve the generation by disentangling various information sources [32]–[34]. In [14], a conditional recurrent neural network (RNN) was considered as the discriminator for GAN where the spatial–temporal information was merged. In [15], the audio and visual features were both used to disentangle the information related to subject and speech from domain knowledge for face reenactment. Such a model was too large and hard to converge due to the adversarial training. This study proposes the flow-based model for mouth reenactment where the conditional likelihood function is maximized to sample the unseen talking mouths in latent space. This flow model continuously transforms the observed data into latent representation via a number of invertible functions where the inverse mapping can be conducted to recover the original observations. The quality of generated data, conditioned on some controllable factor, is accordingly assured [35]. There have been a variety of medical image tasks where the flow model was deployed for vessel segmentation [25] and image transfer from magnetic resonance imaging to positron emission tomography imaging [19]. In [24], a guided invertible domain mapping was proposed for color transfer in conditional image generation. In [36], a Glow-based makeup transfer was developed to estimate a target face image based on the decomposed latent vectors for makeup and face. The disentangled latent representation is crucial.

C. Latent Disentanglement

A key success to conditional generation and mapping is the disentanglement in latent representation which is essential in construction of an unsupervised learning machine. The disentangled representation basically relies on the distinct, separate, modular, and compact factors, learned from observation data, which are independent with minimum information redundancy and interpretable for semantic meaning [37]. In [38], the group theorem was developed as a new perspective to build the disentangled representation. In [39] and [40], a precise criterion with general property was presented to implement the disentangled representation, which connected different symmetry groups in latent space. A disentangling procedure was performed by decomposing each symmetry group into subgroups which preserved the independence. The representation redundancy was minimized to assure model compactness by using the independent generative factors. In [41] and [42], a similar perspective was presented in a way of multidimensional disentangled representation. In [43] and [44], the disentangled representation was performed in an unsupervised manner where the semantic factors were automatically learned from observed data. The model enforced a factorized aggregated posterior, which promoted disentanglement. In [45], a weakly supervised disentanglement was learned with supervision in presence of inductive bias. Recent works have been proposed for flow-based latent disentanglement. In [46], a nonlinear independent component analysis was exploited for disentanglement over a flow model where a Gaussian mixture model was calculated to build a latent space conditioned by classes. In [47], a dedicated neural structure was constructed to separate the mixed images into condition-dependent and independent components. A compact module was learned as a disentangled model driven by a reference condition.

D. Motivation of the Proposed Flows

This article presents the flow-based disentanglement for face reenactment where two learning perspectives are developed. The first perspective is to factorize the latent representation of face images $\mathbf{x}$ into those variables for talking attributes $\mathbf{z}^r$ and general faces $\mathbf{z}^i$. A conditional mouth generation is implemented through the AF flow in accordance with structural and geometric objectives. The second perspective is to carry out the MI flow based on an information-theoretic disentanglement for latent variables $\mathbf{z}^r$ and $\mathbf{z}^i$ where the condition relevant and irrelevant informative objectives are optimized for conditional face generation, respectively. Although two flow models are separately developed, the disentanglement in these two models is consistently performed to find the same attribute or condition relevant and irrelevant variables $\{\mathbf{z}^r, \mathbf{z}^i\}$. Therefore, a cascaded way to combine two flow models under the shared variables can be implemented for mouth reenactment as detailed in the following.

III. AF Flow

First, the AF flow is presented for conditional generation where the observed image $\mathbf{x}$ is transformed to a latent variable $\mathbf{z}$ using a flow model $\mathbf{z} = f_\theta(\mathbf{x})$ with parameter θ. This invertible transformation is to assure the information preserving for data reconstruction. Fig. 2 shows the architecture of AF flow where the target image $\mathbf{x}_\mathcal{T}$ is obtained from a source image $\mathbf{x}_\mathcal{S}$ driven by a query image $\mathbf{x}_q$. Face reenactment aims to generate $\mathbf{x}_\mathcal{T}$ of a target video with the facial features of $\mathbf{x}_q$ whose mouth movement replicates the movement from $\mathbf{x}_\mathcal{S}$ of a source video. AF is performed over latent variables $\mathbf{z}_q$ and $\mathbf{z}_\mathcal{S}$. This study adopts Glow [18], [48] as the backbone, shown by green bars, to carry out the flow-based domain mapping for face images. In the implementation, the squeezed and unsqueezed operations were performed in the begin and end of the input f_θ and output flows f_θ^{-1}, respectively. The multiscale architecture [26] was configured to alleviate the computation cost. A number of objectives are introduced to disentangle $\mathbf{z} = \{\mathbf{z}_q, \mathbf{z}_\mathcal{S}\}$ into an attribute-relevant vector $\mathbf{z}^r = \{\mathbf{z}^r_q, \mathbf{z}^r_\mathcal{S}\}$ for local talking movement of a mouth and an attribute-irrelevant vector $\mathbf{z}^i = \{\mathbf{z}^i_q, \mathbf{z}^i_\mathcal{S}\}$ for global facial structure of a general face. The variable $\mathbf{z}_\mathcal{T}$ is used for face reenactment where the

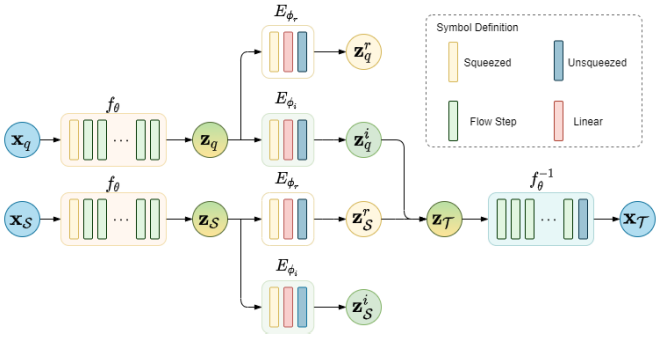

Fig. 2. Architecture for AF flow.

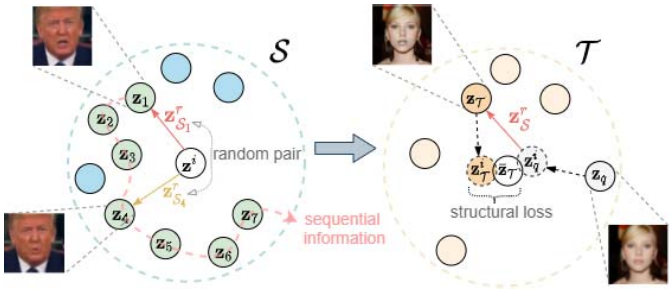

Fig. 3. Right: structural loss between $\mathbf{z}^i_{\mathcal{T}} = E_{\phi_i}(\mathbf{z}^r_{\mathcal{S}} + \mathbf{z}^i_q)$ and $\bar{\mathbf{z}}_{\mathcal{T}}$ in the target domain. Left: Random-pair reconstruction loss for any paired data $\{\mathbf{z}_n, \mathbf{z}_m\}$ in the source domain.

decomposed variables of query vector of facial structure $\mathbf{z}^i_q$ and source vector of lip movement $\mathbf{z}^r_{\mathcal{S}}$ are combined.

A. Factorization by Structure Preserving

In particular, the AF aims to capture the global structure of face samples $\mathbf{x} = \{\mathbf{x}_q, \mathbf{x}_{\mathcal{S}}\}$ from latent vector $\mathbf{z}$, which is common for faces under various talking attributes. Latent vectors $\mathbf{z}^r = E_{\phi_r}(\mathbf{z})$ and $\mathbf{z}^i = E_{\phi_i}(\mathbf{z})$ are extracted by using the attribute-relevant and attribute-irrelevant encoders with parameters ϕ_r and ϕ_i, respectively. This factorization is performed to obtain $\mathbf{z} = \mathbf{z}^r + \mathbf{z}^i$ with the fixed dimension D in $\mathbf{z}$, $\mathbf{z}^r$, and $\mathbf{z}^i$ given by $D = w \times h \times c$ based on the width w, height h, and channel size c of a face image. A video clip of talking face is viewed as a number of face images consisting of the global structure of a real face and the local movement of a facial expression or a lip language. The facial features of a query person in the target domain and the attribute of a talking mouth of a source person are represented by using $\mathbf{z}^i_q$ and $\mathbf{z}^r_{\mathcal{S}}$, respectively, and then merged as $\mathbf{z}_{\mathcal{T}}$ for conditional generation. The attribute-irrelevant encoder E_{ϕ_i} is learned to capture the facial structure by inferring the irrelevant variable $\mathbf{z}^i$ toward the centroid of latent vectors $\mathbf{z}$ of a source person in the source domain $\mathcal{S}$. The structural loss is formed by a square error as the regression loss

$$\mathcal{L}_s = \left\| E_{\phi_i}(\mathbf{z}) - \bar{\mathbf{z}} \right\|^2 \tag{2}$$

where the centroid is calculated as an ensemble mean $\bar{\mathbf{z}} = (1/N)\sum_{n=1}^{N} \mathbf{z}_n$ of the variables $\{\mathbf{z}_n\}_{n=1}^N$ corresponding to the talking frames of a source person. AF flow minimizes this loss to maintain the facial structure of a person in a sequence of talking images. The encoder E_{ϕ_i} is then estimated to preserve the global features of a general face where the attribute-relevant features of talking details are neglected.

It is essential to learn latent disentanglement across various face identities. The disentanglement is strengthened by minimizing the structural loss in the generated target images $\mathbf{x}_{\mathcal{T}}$ that are transferred from source images $\mathbf{x}_{\mathcal{S}}$ conditioned by a query image $\mathbf{x}_q$ where source and query come from different identities. The target sample is generated via $\mathbf{z}_{\mathcal{T}}$ by merging the attribute-relevant feature $\mathbf{z}^r_{\mathcal{S}}$ of source image and the attribute-irrelevant feature $\mathbf{z}^i_q$ of query image. The structural loss in (2) is further measured by using the synthesized feature $\mathbf{z}^r_{\mathcal{S}} + \mathbf{z}^i_q$ in the target domain via

$$\mathcal{L}_t = \left\| E_{\phi_i}\left(E_{\phi_r}(\mathbf{z}_{\mathcal{S}}) + \mathbf{z}^i_q\right) - \bar{\mathbf{z}}_{\mathcal{T}} \right\|^2 \tag{3}$$

where $\bar{\mathbf{z}}_{\mathcal{T}}$ denotes the ensemble mean of target images. This loss is minimized to preserve the structure of target samples due to domain mapping $\mathcal{S} \rightarrow \mathcal{T}$. Fig. 3 (right) shows the structural loss measured by the synthesized variable $\mathbf{z}_{\mathcal{T}}$ in the target domain by adding the latent variables of the query face $\mathbf{z}^i_q$ and the talking information of source person $\mathbf{z}^r_{\mathcal{S}}$ (encoded by E_{ϕ_r} and depicted by red). The structure encoder E_{ϕ_i} and movement encoder E_{ϕ_r} are optimized to preserve the structures before and after domain mapping.

B. Factorization by Self-Supervised Learning

In practice, the stereo paired data between two domains are missing in the task of face reenactment. Supervised factorization is not possible. However, the relation between a pair of any two talking frames within the same video clip of face images can be characterized in a self-supervised way. This study presents the random-pair reconstruction loss $\mathcal{L}_p$ that is minimized to estimate E_{ϕ_i} and E_{ϕ_r}. As shown in Fig. 3 (left), the image frames of a single sequence are self-collected to form a set of pseudo-paired data $\{\mathbf{z}_n, \mathbf{z}_m\}$ where two different frames within a single sequence from a source person are randomly selected. The reconstruction error due to a latent variable $\mathbf{z}_m$ is then calculated by

$$\mathcal{L}_p = \left\| E_{\phi_i}(\mathbf{z}_n) + E_{\phi_r}(\mathbf{z}_m) - \mathbf{z}_m \right\|^2 \quad \forall n, m \in \{1, 2, \ldots, N\}; \quad n \neq m \tag{4}$$

which is minimized to preserve the structure information and attribute evidence. This method does not only augment the positive pairs but also encourage the training diversity.

The previous three losses $\{\mathcal{L}_s, \mathcal{L}_t, \mathcal{L}_p\}$ closely affect the structure encoder E_{ϕ_i}. To enhance the flow-based disentanglement, AF flow is further consolidated by minimizing the classification loss due to a word label k of talking mouth along a training sequence that is predicted by using the talking variable $(\mathbf{z}^r)_k$ [49]. This variable is considerably affected by the attribute encoder E_{ϕ_r}. The cross-entropy error between one-hot class output $\mathbf{y}_n = \{y_{nk}\}$ and posterior output C_ψ of a neural sequential classifier with parameter ψ is measured by

$$\mathcal{L}_y = -\sum_{n=1}^{N}\sum_{k=1}^{N_y} y_{nk} \log\left(C_\psi\left(\left(\mathbf{z}^r_n\right)_k\right)\right) \tag{5}$$

where N_y denotes the vocabulary size. This classification loss due to word label is minimized to jointly train two encoders $\{E_{\phi_i}, E_{\phi_r}\}$ and one classifier C_ψ for AF.

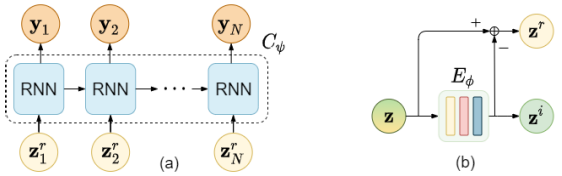

Fig. 4. Architectures for (a) sequential classifier with vanilla RNN and (b) single encoder for factorization of attributes.

C. Implementation and Optimization

In this study, neural sequential classifier C_ψ with word outputs $\{\mathbf{y}_n\}_{n=1}^N$ is implemented by an RNN using a sequence of attribute-relevant vectors $\{\mathbf{z}_n^r\}_{n=1}^N$ as the inputs, which is shown in Fig. 4(a). In addition, a single encoder E_{ϕ_i} or E_ϕ is configured as in Fig. 4(b) instead of using two encoders $\{E_{\phi_r}, E_{\phi_i}\}$ in Fig. 2. This scheme simplifies the training convergence and makes sure the inverse procedure through the additive relation $\mathbf{z} = \mathbf{z}^r + \mathbf{z}^i$. There are two training stages for AF flow. The first stage is to train an unsupervised Glow model $\mathbf{z} = f_\theta(\mathbf{x})$ with parameter θ by maximizing the likelihood of observed data $\mathbf{x}$ or minimizing the flow loss in (1). However, the observed data $\mathbf{x}$ consist of discrete pixel values of images. In the implementation, the dequantization method [18], [27] is applied by using a noise random variable $\boldsymbol{\varepsilon}$ in a flow model. The flow transformation and its invertible function are obtained by $\mathbf{z} = f_\theta(\widetilde{\mathbf{x}})$ and $\widetilde{\mathbf{x}} = f_\theta^{-1}(\mathbf{z}; \boldsymbol{\varepsilon})$, respectively, where $\widetilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\varepsilon}$ is obtained by adding the positive uniform sample of a noise signal drawn by $\boldsymbol{\varepsilon} \sim \mathcal{U}(0, b)$ with a small value of bounding parameter b. This Glow model using θ is then adopted to collect the mini-batches of training sequences $\{\mathbf{z}, \mathbf{y}\}$. The second stage is to use them to fulfill the AF by continuously updating two encoders and one classifier with parameters $\{\phi, \psi\}$ where the combined AF loss is obtained by

$$\mathcal{L}_{\text{AF}}(\mathbf{z}, \mathbf{y}; \phi, \psi) = \mathcal{L}_s(\mathbf{z}; \phi) + \mathcal{L}_t(\mathbf{z}; \phi) + \mathcal{L}_p(\mathbf{z}; \phi) + \mathcal{L}_y(\mathbf{z}, \mathbf{y}; \phi, \psi) \quad (6)$$

which is minimized by calculating the gradients $(\partial \mathcal{L}_{\text{AF}}/\partial \phi)$ and $(\partial \mathcal{L}_{\text{AF}}/\partial \psi)$ in the stochastic gradient descent (SGD) algorithm. AF flow carries out the disentanglement for conditional generation via the structure encoder E_ϕ and classifier C_ψ by minimizing the structural losses in two domains: the random pair loss and the classification loss. In what follows, an alternative disentanglement with information evidence is presented. Algorithm 1 shows the learning stages of parameters $\{\theta, \phi, \psi\}$ in AF flow.

Algorithm 1 Learning Procedure for AF Flow Model

Input queries, source samples and labels $\{\mathbf{x}_q, \mathbf{x}_\mathcal{S}, \mathbf{y}\}$
Initialize parameters θ, ϕ_r, ϕ_i, ψ and select parameter b
while θ, ϕ_r, ϕ_i, ψ *not converged* **do**
  update θ using $\mathbf{x}^j$ from $\{\mathbf{x}_q, \mathbf{x}_\mathcal{S}\}$ via gradient of reconstruction error using $\widetilde{\mathbf{x}}^j \leftarrow \mathbf{x}^j + \boldsymbol{\varepsilon}$
  for *each mini-batch* $\mathbf{x}^j$, $\mathbf{y}^j$ *from* $\{\mathbf{x}_q, \mathbf{x}_\mathcal{S}, \mathbf{y}\}$ **do**
    calculate $\mathbf{z}$, $\mathbf{z}^r$, $\mathbf{z}^i$ via f_θ, E_{ϕ_r}, E_{ϕ_i} using $\mathbf{x}^j$
    calculate $\bar{\mathbf{z}}$, source structural loss $\mathcal{L}_s$ in (2)
    calculate $\mathbf{z}_\mathcal{T}$, $\bar{\mathbf{z}}_\mathcal{T}$, target structural loss $\mathcal{L}_t$ in (3)
    calculate random pairs $\{\mathbf{z}_n, \mathbf{z}_m\}$ and reconstruction loss $\mathcal{L}_p$ in (4)
    calculate classifier loss $\mathcal{L}_y$ in (5) using $\mathbf{y}^j$, C_ψ
    calculate AF loss $\mathcal{L}_{\text{AF}}$ in (6)
    update ϕ_r, ϕ_i, ψ via gradients $((\partial \mathcal{L}_{\text{AF}})/(\partial \phi_r))$, $((\partial \mathcal{L}_{\text{AF}})/(\partial \phi_i))$, $((\partial \mathcal{L}_{\text{AF}})/(\partial \psi))$
  end
end

IV. MI Flow

This study further presents the MI flow for latent disentanglement where the attribute relevance and irrelevance are factorized.

A. Information-Theoretic Disentanglement

A key property of disentanglement is to factorize latent variables with distinct features. Consider the disentanglement of latent variable $\mathbf{z}$ into condition-irrelevant variable $\mathbf{z}^i$ and condition-relevant variable $\mathbf{z}^r$, where $\mathbf{z} = \{\mathbf{z}^i, \mathbf{z}^r\}$, the learning

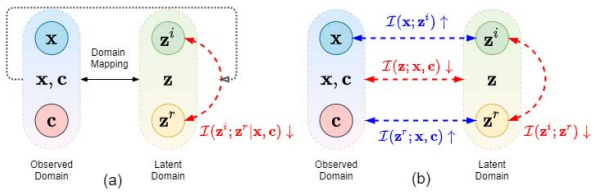

Fig. 5. Illustration for (a) MI in latent domain and (b) informative mapping and disentanglement based on the condition irrelevant and relevant variables $\{\mathbf{z}^i, \mathbf{z}^r\}$.

objective is formed as an MI loss $\mathcal{L} = \mathcal{I}(\mathbf{z}^i; \mathbf{z}^r | \mathbf{x}, \mathbf{c})$ which is minimized by using the observed data $\mathbf{x}$ (corresponding to a source sample $\mathbf{x}_\mathcal{S}$) and a relevance condition $\mathbf{c}$ (corresponding to his/her subimage of mouth region). Mapping between observed domain and latent domain is characterized. Fig. 5(a) shows how the separation between $\mathbf{z}^i$ (for facial structure) and $\mathbf{z}^r$ (for lip movement) is increased by minimizing $\mathcal{I}(\mathbf{z}^i; \mathbf{z}^r | \mathbf{x}, \mathbf{c})$. However, direct calculation of true MI is difficult. This MI is therefore arranged by manipulating the entropy terms $\mathcal{H}(\cdot)$ that are factorized in a form of

$$\begin{aligned}
&\mathcal{I}(\mathbf{z}^i; \mathbf{z}^r | \mathbf{x}, \mathbf{c}) \\
&= \mathcal{H}(\mathbf{z}^i | \mathbf{x}) + \mathcal{H}(\mathbf{z}^r | \mathbf{x}, \mathbf{c}) - \mathcal{H}(\mathbf{z}^i, \mathbf{z}^r | \mathbf{x}, \mathbf{c}) \\
&= -\mathcal{H}(\mathbf{z}^i) + \mathcal{H}(\mathbf{z}^i | \mathbf{x}) - \mathcal{H}(\mathbf{z}^r) + \mathcal{H}(\mathbf{z}^r | \mathbf{x}, \mathbf{c}) + \mathcal{H}(\mathbf{z}^i, \mathbf{z}^r) \\
&\quad -\mathcal{H}(\mathbf{z}^i, \mathbf{z}^r | \mathbf{x}, \mathbf{c}) + \mathcal{H}(\mathbf{z}^i) + \mathcal{H}(\mathbf{z}^r) - \mathcal{H}(\mathbf{z}^i, \mathbf{z}^r) \\
&= -\mathcal{I}(\mathbf{z}^i; \mathbf{x}) - \mathcal{I}(\mathbf{z}^r; \mathbf{x}, \mathbf{c}) + \mathcal{I}(\mathbf{z}; \mathbf{x}, \mathbf{c}) + \mathcal{I}(\mathbf{z}^i; \mathbf{z}^r) \\
&= \underbrace{\mathcal{I}(\mathbf{z}; \mathbf{x}, \mathbf{c})}_{\text{domain mapping}} \quad \underbrace{-\mathcal{I}(\mathbf{z}^i; \mathbf{x})}_{\text{condition irrelevance}} \quad \underbrace{-\mathcal{I}(\mathbf{z}^r; \mathbf{x}, \mathbf{c}) + \mathcal{I}(\mathbf{z}^i; \mathbf{z}^r)}_{\text{condition relevance}}
\end{aligned} \quad (7)$$

which consists of four individual MI terms as shown in Fig. 5(b). The first term is a domain mapping MI $\mathcal{I}(\mathbf{z}; \mathbf{x}, \mathbf{c})$, which is minimized to disentangle the relation between observed domain $\{\mathbf{x}, \mathbf{c}\}$ and latent domain $\mathbf{z}$. The second term $\mathcal{I}(\mathbf{z}^i; \mathbf{x})$ denotes the condition-irrelevant MI, which is maximized to infer the condition-irrelevant variable $\mathbf{z}^i$ so as to sufficiently reflect the observed data $\mathbf{x}$ where the

condition $\mathbf{c}$ is missing. The last two terms reflect the condition-relevant MI. The MI $\mathcal{I}(\mathbf{z}^r;\mathbf{x},\mathbf{c})$ is maximized to pursue the condition-relevant variable $\mathbf{z}^r$, which is substantially correlated with data and condition $\{\mathbf{x},\mathbf{c}\}$. The MI $\mathcal{I}(\mathbf{z}^i;\mathbf{z}^r)$ is minimized to disentangle the condition-irrelevant and relevant variables $\{\mathbf{z}^i,\mathbf{z}^r\}$. In [50], the information bottleneck in the invertible neural network was proposed to represent the information constraint for a generative classifier to optimally balance between classification accuracy and model complexity. This study presents the MI-based flow model where four MI terms are jointly optimized for disentanglement of structure and attribute variables $\{\mathbf{z}^i,\mathbf{z}^r\}$.

B. MI Objectives

The first MI objective $\mathcal{I}(\mathbf{z};\mathbf{x},\mathbf{c})$ is used to disentangle domain mapping between $\{\mathbf{x},\mathbf{c}\}$ and $\mathbf{z}$, which is driven by the posterior distribution $p_\theta(\mathbf{z}|\mathbf{x},\mathbf{c})$ using parameter θ. Flow model is applied to transform from observed data $\mathbf{x}$ to latent variable $\mathbf{z}$ by using invertible function $\mathbf{z}=f_\theta(\mathbf{x})$ with parameter θ, while the condition $\mathbf{c}$ provides the attribute as the prior information for conditional generation. In particular, minimizing this domain mapping MI is equivalent to minimizing its variational upper bound expressed as [51]

$$\mathcal{I}(\mathbf{z};\mathbf{x},\mathbf{c})\le\mathbb{E}_{p(\mathbf{x})}[D_{\mathrm{KL}}(p_\theta(\mathbf{z}|\mathbf{x},\mathbf{c})\|p(\mathbf{z}|\mathbf{c}))]+\mathcal{H}(\mathbf{c}) \tag{8}$$

where the entropy $\mathcal{H}(\mathbf{c})$ is independent of flow parameter θ. Optimization problem turns out to minimizing the Kullback–Leiblier (KL) divergence $\mathbb{E}_{p(\mathbf{x})}[D_{\mathrm{KL}}(p_\theta(\mathbf{z}|\mathbf{x},\mathbf{c})\|p(\mathbf{z}|\mathbf{c}))]$ for latent disentanglement. Considering the property of inverse function in two transformation directions and the dequantization scheme mentioned in Section III-C, we obtain the parametric distributions $p_\theta(\mathbf{z}|\mathbf{x},\mathbf{c})=p(f_\theta^{-1}(\mathbf{z};\boldsymbol{\varepsilon})|\mathbf{x},\mathbf{c})\left|\det((df_\theta^{-1}(\mathbf{z};\boldsymbol{\varepsilon}))/d\mathbf{z})\right|$ and $p_\theta(\mathbf{x}|\mathbf{c})=p(f_\theta(\mathbf{x})|\mathbf{c})\left|\det((df_\theta(\mathbf{x}))/d\mathbf{x})\right|$, which are applied to reformulate the KL term in (8) as

$$\begin{aligned}
&\mathbb{E}_{p(\mathbf{x})}[D_{\mathrm{KL}}(p_\theta(\mathbf{z}|\mathbf{x},\mathbf{c})\|p(\mathbf{z}|\mathbf{c}))]\\
&=\int p(\mathbf{x})\left[\int p_\theta(\mathbf{z}|\mathbf{x},\mathbf{c})\log\frac{p_\theta(\mathbf{z}|\mathbf{x},\mathbf{c})}{p(\mathbf{z}|\mathbf{c})}d\mathbf{z}\right]d\mathbf{x}\\
&=\int p(\mathbf{x})\left[\int p\big(f_\theta^{-1}(\mathbf{z};\boldsymbol{\varepsilon})|\mathbf{x},\mathbf{c}\big)\left|\det\frac{df_\theta^{-1}(\mathbf{z};\boldsymbol{\varepsilon})}{d\mathbf{z}}\right|\right.\\
&\qquad\left.\times\log\frac{p\big(f_\theta^{-1}(\mathbf{z};\boldsymbol{\varepsilon})|\mathbf{x},\mathbf{c}\big)\left|\det\frac{df_\theta^{-1}(\mathbf{z};\boldsymbol{\varepsilon})}{d\mathbf{z}}\right|}{p(f_\theta(\widetilde{\mathbf{x}})|\mathbf{c})}d\mathbf{z}\right]d\mathbf{x}\\
&=\int p\big(f_\theta^{-1}(\mathbf{z};\boldsymbol{\varepsilon})|\mathbf{c}\big)\left|\det\frac{df_\theta^{-1}(\mathbf{z};\varepsilon)}{d\mathbf{z}}\right|\\
&\quad\times\log\frac{p\big(f_\theta^{-1}(\mathbf{z};\boldsymbol{\varepsilon})|\mathbf{c}\big)}{p(f_\theta(\widetilde{\mathbf{x}})|\mathbf{c})\left|\det\frac{df_\theta(\widetilde{\mathbf{x}})}{d\widetilde{\mathbf{x}}}\right|}d\mathbf{z}=\int p(\widetilde{\mathbf{x}}|\mathbf{c})\log\frac{p(\widetilde{\mathbf{x}}|\mathbf{c})}{p_\theta(\widetilde{\mathbf{x}}|\mathbf{c})}d\widetilde{\mathbf{x}}\\
&=D_{\mathrm{KL}}(p(\widetilde{\mathbf{x}}|\mathbf{c})\|p_\theta(\widetilde{\mathbf{x}}|\mathbf{c}))=\mathcal{H}(\widetilde{\mathbf{x}}|\mathbf{c})-\mathbb{E}_{p(\widetilde{\mathbf{x}}|\mathbf{c})}[\log p_\theta(\widetilde{\mathbf{x}}|\mathbf{c})].
\end{aligned}\tag{9}$$

Since $\mathcal{H}(\widetilde{\mathbf{x}}|\mathbf{c})$ is independent of flow parameter θ, the MI loss for domain mapping is accordingly obtained by

$$\mathcal{L}_d=-\mathbb{E}_{p(\widetilde{\mathbf{x}}|\mathbf{c})}[\log p_\theta(\widetilde{\mathbf{x}}|\mathbf{c})] \tag{10}$$

which is minimized to learn the flow-based representation. Equivalently, the conditional likelihood of noisy samples $\widetilde{\mathbf{x}}$ with condition $\mathbf{c}$ is maximized to learn the flow model.

Next, the condition-irrelevant MI $\mathcal{I}(\mathbf{x};\mathbf{z}^i)$ is maximized to infer $\mathbf{z}^i$, which sufficiently reflects $\mathbf{x}$ but irrelevantly relates to attribute $\mathbf{c}$. This MI is arranged to find a lower bound via

$$\begin{aligned}
\mathcal{I}(\mathbf{x};\mathbf{z}^i)&=\mathcal{H}(\mathbf{x})-\mathcal{H}(\mathbf{x}|\mathbf{z}^i)\\
&=\mathbb{E}_{p(\mathbf{z}^i|\mathbf{x})}\big[\mathbb{E}_{\mathbf{x}\sim g_\theta(\mathbf{z}^i)}[\log p(\mathbf{x}|\mathbf{z}^i)]\big]+\mathcal{H}(\mathbf{x})\\
&=\mathbb{E}_{p(\mathbf{z}^i|\mathbf{x})}\big[D_{\mathrm{KL}}(p(\mathbf{x}|\mathbf{z}^i)\|p_\theta(\mathbf{x}|\mathbf{z}^i))\\
&\qquad+\mathbb{E}_{\mathbf{x}\sim g_\theta(\mathbf{z}^i)}[\log p_\theta(\mathbf{x}|\mathbf{z}^i)]\big]+\mathcal{H}(\mathbf{x})\\
&\ge\mathbb{E}_{p(\mathbf{z}^i|\mathbf{x})}\big[\mathbb{E}_{\mathbf{x}\sim g_\theta(\mathbf{z}^i)}[\log p_\theta(\mathbf{x}|\mathbf{z}^i)]\big]+\mathcal{H}(\mathbf{x})
\end{aligned}\tag{11}$$

where $\mathbf{x}$ is sampled from inverse function g_θ or f_θ^{-1} using $\mathbf{z}^i$ and the auxiliary distribution $p_\theta(\mathbf{x}|\mathbf{z}^i)$ is merged to approximate true posterior $p(\mathbf{x}|\mathbf{z}^i)$. The lower bound in (11) is obtained since the KL term is always nonnegative. Notably, rather than using an additional decoder to approximate true distribution, it is meaningful to reuse flow model by reversing its transformation direction or equivalently applying its inverse function to implement this generator or decoder. Flow parameter θ is not only affected by domain mapping objective $\mathcal{I}(\mathbf{z};\mathbf{x},\mathbf{c})$ but also by condition-irrelevant objective $\mathcal{I}(\mathbf{x};\mathbf{z}^i)$. Such a scheme is helpful to train a flow-based generator $p_\theta(\mathbf{x}|\mathbf{z}^i)$. Maximizing $\mathcal{I}(\mathbf{x};\mathbf{z}^i)$ is comparable with maximizing its lower bound. As a result, the loss function for condition-irrelevant MI is constructed by removing the independent term $\mathcal{H}(\mathbf{x})$ to form the objective

$$\mathcal{L}_i=-\mathbb{E}_{p(\mathbf{z}^i|\mathbf{x})}\big[\mathbb{E}_{\mathbf{x}\sim g_\theta(\mathbf{z}^i)}[\log p_\theta(\mathbf{x}|\mathbf{z}^i)]\big]. \tag{12}$$

In addition, the informative latent disentanglement is further strengthened by inferring the condition-relevant latent variable $\mathbf{z}^r$ where the correlation between the given condition $\mathbf{c}$ and the disentangled embedding $\mathbf{z}^r$ is increased by optimizing the condition-relevant MI $-\mathcal{I}(\mathbf{z}^r;\mathbf{x},\mathbf{c})+\mathcal{I}(\mathbf{z}^i;\mathbf{z}^r)$. There are two terms in learning objective. The first term $\mathcal{I}(\mathbf{z}^r;\mathbf{x},\mathbf{c})$ is maximized to consolidate the condition-relevant variable $\mathbf{z}^r$, which reflects the image $\mathbf{x}$ as well as the condition $\mathbf{c}$. Similarly, this term can be factorized and manipulated as

$$\begin{aligned}
\mathcal{I}(\mathbf{z}^r;\mathbf{x},\mathbf{c})&=\mathcal{H}(\mathbf{z}^r)-\mathcal{H}(\mathbf{z}^r|\mathbf{x},\mathbf{c})\\
&=\mathbb{E}_{p(\mathbf{x},\mathbf{c})}\big[\mathbb{E}_{p_\theta(\mathbf{z}^r|\mathbf{x},\mathbf{c})}[\log p_\theta(\mathbf{z}^r|\mathbf{x},\mathbf{c})]\big]+\mathcal{H}(\mathbf{z}^r)\\
&=\mathbb{E}_{p(\mathbf{x},\mathbf{c})}\big[\mathbb{E}_{p_\theta(\mathbf{z}^r|\mathbf{x},\mathbf{c})}[\log p_\varphi(\mathbf{z}^r|\mathbf{x},\mathbf{c})]\big]\\
&\quad+\mathbb{E}_{p(\mathbf{x},\mathbf{c})}\big[D_{\mathrm{KL}}(p_\theta(\mathbf{z}^r|\mathbf{x},\mathbf{c})\|p_\varphi(\mathbf{z}^r|\mathbf{x},\mathbf{c}))\big]+\mathcal{H}(\mathbf{z}^r)\\
&\ge\mathbb{E}_{p(\mathbf{x},\mathbf{c})}\big[\mathbb{E}_{p_\theta(\mathbf{z}^r|\mathbf{x},\mathbf{c})}[\log p_\varphi(\mathbf{z}^r|\mathbf{x},\mathbf{c})]\big]+\mathcal{H}(\mathbf{z}^r).
\end{aligned}\tag{13}$$

Again, this lower bound is obtained due to the nonnegative KL term. Notably, in (13), a learnable conditional distribution $p_\varphi(\mathbf{z}^r|\mathbf{x},\mathbf{c})$ or $p_\varphi(\mathbf{z}^r|\mathbf{c})$ with Gaussian parameter φ with mean μ and standard deviation σ is incorporated to provide prior information for condition-relevant variable $\mathbf{z}^r$, which is calculated by flow model f_θ under a distribution $p_\theta(\mathbf{z}^r|\mathbf{x},\mathbf{c})$. Alternatively, the second term is minimized to disentangle $\mathbf{z}^r$ from $\mathbf{z}^i$. This term can be factorized as $\mathcal{I}(\mathbf{z}^i;\mathbf{z}^r)=\mathcal{H}(\mathbf{z}^r)-\mathcal{H}(\mathbf{z}^r|\mathbf{z}^i)$ and combined with the first term in (13) to derive

the variational upper bound of condition-relevant MI

$$\begin{aligned}-\mathcal{I}(\mathbf{z}^r;\mathbf{x},\mathbf{c})+\mathcal{I}(\mathbf{z}^i;\mathbf{z}^r)&=\mathcal{H}(\mathbf{z}^r|\mathbf{x},\mathbf{c})-\mathcal{H}(\mathbf{z}^r|\mathbf{z}^i)\\&\le-\mathbb{E}_{p(\mathbf{x},\mathbf{c})}\big[\mathbb{E}_{p_\theta(\mathbf{z}^r|\mathbf{x},\mathbf{c})}[\log p_\varphi(\mathbf{z}^r|\mathbf{x},\mathbf{c})]\big]-\mathcal{H}(\mathbf{z}^r|\mathbf{z}^i).\end{aligned}\tag{14}$$

Minimizing this MI objective is equivalent to minimizing its corresponding upper bound. The loss function due to condition-relevant MI is then obtained by minimizing

$$\mathcal{L}_r=-\mathbb{E}_{p(\mathbf{x},\mathbf{c})}\big[\mathbb{E}_{p_\theta(\mathbf{z}^r|\mathbf{x},\mathbf{c})}[\log p_\varphi(\mathbf{z}^r|\mathbf{x},\mathbf{c})]\big]\tag{15}$$

to update flow parameter θ as well as prior parameter φ.

C. Learning Algorithm

In the implementation, latent disentanglement using domain mapping MI can be further strengthened by expanding the loss function $\mathcal{L}_d$. The generative likelihood given by condition $\mathbf{c}$ in Eq. (10) is extended by jointly considering condition-irrelevant prior $p(\mathbf{z}^i)$ and condition-relevant prior $p_\varphi(\mathbf{z}^r|\mathbf{c})$ based on $\mathbf{z}=\{\mathbf{z}^i,\mathbf{z}^r\}$. The parameters of flow model and prior model $\{\theta,\varphi\}$ are merged in the derivation as

$$\begin{aligned}\mathcal{L}_d&=-\mathbb{E}_{p(\mathbf{x}|\mathbf{c})}[\log p_\theta(\mathbf{x}|\mathbf{c})]\\&=-\mathbb{E}_{p(\mathbf{x}|\mathbf{c})}\left[\log p_\varphi(f_\theta(\mathbf{x})|\mathbf{c})+\log\left|\det\frac{df_\theta(\mathbf{x})}{d\mathbf{x}}\right|\right]\\&=-\mathbb{E}_{p(\mathbf{x}|\mathbf{c})}\left[\mathbb{E}_{p_\theta(\mathbf{z}|\mathbf{x},\mathbf{c})}\left[\log p(\mathbf{z}^i)+\log p_\varphi(\mathbf{z}^r|\mathbf{c})+\log\left|\det\frac{d\mathbf{z}}{d\mathbf{x}}\right|\right]\right]\\&=-\mathbb{E}_{p(\mathbf{x}|\mathbf{c})}\left[\mathbb{E}_{p_\theta(\mathbf{z}^i|\mathbf{x},\mathbf{c})}\big[\log p(\mathbf{z}^i)\big]+\mathbb{E}_{p_\theta(\mathbf{z}|\mathbf{x},\mathbf{c})}\left[\log\left|\det\frac{d\mathbf{z}}{d\mathbf{x}}\right|\right]\right]\\&\quad\underbrace{-\mathbb{E}_{p(\mathbf{x}|\mathbf{c})}\big[\mathbb{E}_{p_\theta(\mathbf{z}^r|\mathbf{x},\mathbf{c})}\big[\log p_\varphi(\mathbf{z}^r|\mathbf{c})\big]\big]}_{\mathcal{L}_r}.\end{aligned}\tag{16}$$

The loss function for condition-relevant MI $\mathcal{L}_r$ in (15) is seen as a part of loss term for domain mapping MI $\mathcal{L}_d$. Optimizing the conditional prior model $p_\varphi(\mathbf{z}^r|\mathbf{c})$ or equivalently $p_\varphi(\mathbf{z}^r|\mathbf{x},\mathbf{c})$ is performed under the training of the conditional flow model f_θ, which is used in $p_\theta(\mathbf{z}|\mathbf{x},\mathbf{c})$. In the optimization, this conditional prior is provided to infer the condition-relevant variable $\mathbf{z}^r$, while the condition-irrelevant variable $\mathbf{z}^i$ simply relies on a standard Gaussian prior. Besides, the loss functions $\mathcal{L}_d$ and $\mathcal{L}_r$ depend on input data $\{\mathbf{x},\mathbf{c}\}$ as well as model parameters $\{\theta,\varphi\}$. The loss function for condition-irrelevant MI $\mathcal{L}_i(\mathbf{x};\theta)$ in (12) is only related to observed data $\mathbf{x}$ and flow parameter θ. Assuming that the likelihood function $p_\theta(\mathbf{x}|\mathbf{z}^i)$ given by condition-irrelevant variable $\mathbf{z}^i$, calculated by flow model f_θ, is Gaussian with zero mean for reconstruction error and unit variance in each dimension. Loss function $\mathcal{L}_i$ is then simply seen as the aggregation of reconstruction errors from individual observations $\mathbf{x}$ as

$$\mathcal{L}_i(\mathbf{x};\theta)=\mathbb{E}_{p(\mathbf{z}^i|\mathbf{x})}\big[\mathbb{E}_{\mathbf{x}\sim g_\theta(\mathbf{z}^i)}[\|\mathbf{x}-g_\theta(\mathbf{z}^i)\|^2]\big].\tag{17}$$

This loss is minimized to build a flow model where its inverse g_θ works toward the smallest reconstruction error. Such an inverse flow model is reused to act as the conditional generator. Note that only the condition-irrelevant variable $\mathbf{z}^i$ is used as the input to generator. This property makes sure of an informative latent variable $\mathbf{z}^r$, which substantially reflects its relation with the observed input $\mathbf{x}$. Therefore, the loss function of MI flow is combined with a parameter α as

$$\mathcal{L}_{\mathrm{MI}}(\mathbf{x},\mathbf{c};\theta,\varphi)=\mathcal{L}_d(\mathbf{x},\mathbf{c};\theta,\varphi)+\alpha\mathcal{L}_r(\mathbf{x},\mathbf{c};\theta,\varphi)+\mathcal{L}_i(\mathbf{x};\theta).\tag{18}$$

Algorithm 2 Learning Procedure for MI Flow Model

```
Input training mini-batches x = {x^j} and condition c
Initialize parameters θ, φ and select parameters b, α
while θ, φ not converged do
    for each mini-batch x^j do
        select c for mini-batch x^j
        find a noise sample by ε ~ U(0, b)
        de-quantize x^j by x̃^j ← x^j + ε
        calculate z^r, z^i, z via f_θ using x̃^j
        calculate conditional prior p_φ(z^r|c)
        calculate posteriors p_θ(z^r|x, c), p_θ(z^i|x, c)
        generate samples via x̂^j ~ g_θ(z^i)
        calculate relevance loss L_r in (15)
        calculate domain loss L_d in (16)
        calculate irrelevance loss L_i in (17)
        calculate MI loss L_MI in (18)
        update θ, φ via gradients
        ((∂L_MI)/(∂θ)),((∂L_MI)/(∂φ))
    end
end
```

Algorithm 2 illustrates the learning procedure of the proposed MI flow where the flow model f_θ or g_θ with parameter θ and the Gaussian prior $p_\varphi(\mathbf{z}^r|\mathbf{c})$ with parameters $\varphi=\{\mu,\sigma\}$ are tightly merged and jointly optimized. This is different from the separate two-stage training in AF flow.

D. Architecture and Implementation

Extended from the concept in Fig. 5, the architecture of the proposed MI flow is configured in Fig. 6(a) as the training stage and Fig. 6(b) as the generation stage where Fig. 6(c) describes the definition of different symbols. Different from AF flow using two encoders E_{ϕ_r} and E_{ϕ_i} for factorization of flow-based latent vector $\mathbf{z}=f_\theta(\mathbf{x})$ into $\mathbf{z}^r$ and $\mathbf{z}^i$, respectively, the flow model in MI flow is factorized as $f_\theta=f_\theta^r\circ f_\theta^i$, which is applied to calculate the condition relevant and irrelevant vectors $\mathbf{z}=\{\mathbf{z}^r,\mathbf{z}^i\}$ from the observed vector $\mathbf{x}$ by using $\{f_\theta^r,f_\theta^i\}$ via

$$\mathbf{h}=f_\theta^r(\mathbf{x}),\quad\{\mathbf{z}^r,\mathbf{h}^i\}=\mathrm{split}(\mathbf{h}),\ \mathbf{z}^i=f_\theta^i(\mathbf{h}^i).\tag{19}$$

Function split$(\cdot)$ is used for variable splitting. Such a factorization is fulfilled to implement latent disentanglement $\mathbf{z}=\{\mathbf{z}^r,\mathbf{z}^i\}=f_\theta(\mathbf{x})$ with the inverse $\mathbf{x}=f_\theta^{-1}(\mathbf{z}^r,\mathbf{z}^i)$. In the training stage, the flow components f_θ^r and f_θ^i consisted of K flow steps or coupling layers. Considering the image data with three channels $64\times64\times3$, the flow component f_θ^r first adopted a squeezed operation to increase the channel number and reshape three-way tensor input as $32\times32\times12$.

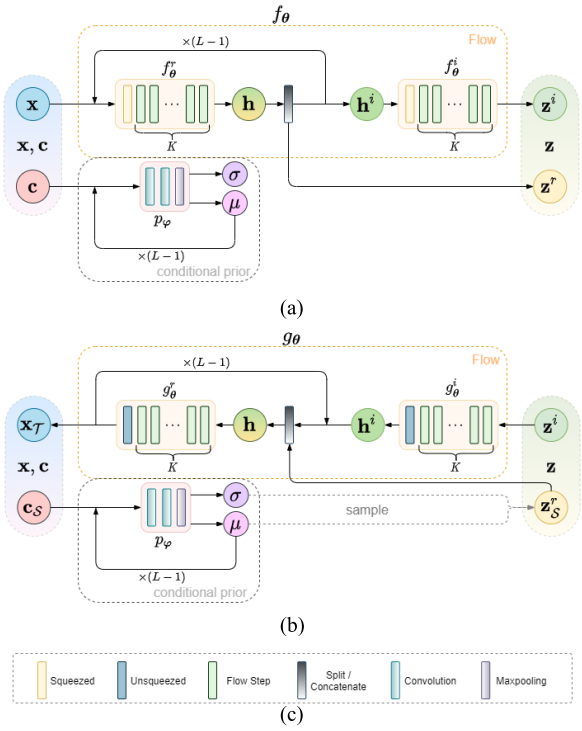

Fig. 6. Architectures for MI flow in training and generation stages. Different symbols are defined (a) Training stage. (b) Generation stage. (c) Symbol definition.

The remaining layers had the same size. After these flow steps, the output $\mathbf{h}$ was split into two variables with the same shape $32 \times 32 \times 6$ where one variable was used as the condition-relevant variable $\mathbf{z}^r$ and the other variable $\mathbf{h}^i$ was used as the input to repeat this flow step. Such a computation block was repeated $L-1$ times to construct a multiscale layer architecture [26] so as to produce $\mathbf{z}^i$. The condition-irrelevant vector $\mathbf{z}^i$ was obtained after the flow component f^i_θ with K flow steps was computed. This architecture was useful to reduce the computation cost and improve the model regularization. Notably, this Glow model is not only used in MI flow but also employed in AF flow when finding $\mathbf{z}_q$ and $\mathbf{z}_\mathcal{S}$ in Fig. 2. Different from AF flow, the conditional prior is incorporated in MI flow to draw the sample of condition-relevant variable $\mathbf{z}^r$ based on the Gaussian mean and variance parameters $\{\mu, \sigma\}$ where $L-1$ layers of convolution and max pooling were calculated to find $p_\varphi(\mathbf{z}^r|\mathbf{c})$ by given the condition image $\mathbf{c}$. This process was to match the size of multiscale architecture. Parameter φ of conditional prior network was estimated to tightly impose the condition of lip movement to infer the relevance vector $\mathbf{z}^r$. Here, $K = 32$ and $L = 4$ were used.

After the training stage, the generation stage was implemented for face reenactment by inverting all of computations by using the inverse functions $\{g^i_\theta, g^r_\theta\}$ and changing the splitting as the concatenating where the condition-irrelevant and condition-relevant clues from $\{\mathbf{z}^i, \mathbf{z}^r\}$ were used in each computation block. There were $L-1$ blocks. Given a condition of lip movement $\mathbf{c}_\mathcal{S}$ of a query image $\mathbf{x}_q$ in the source domain, the proposed MI flow is able to generate a sequence of images of a target face $\mathbf{x}_\mathcal{T}$. This model implements the conditional generation for domain mapping between $\mathbf{c}_\mathcal{S}$ and $\mathbf{x}_\mathcal{T}$. In the implementation, the trained conditional prior network $p_\varphi(\mathbf{z}^r|\mathbf{c})$ is applied to draw the sample of condition-relevant variable $\mathbf{z}^r_\mathcal{S}$ by using query condition $\mathbf{c}_\mathcal{S}$, which is then concatenated with $\mathbf{h}^i$. This $\mathbf{h}^i$ is calculated by the inverse model g^i_θ, due to the condition-irrelevant variable $\mathbf{z}^i$. The concatenated vector $\mathbf{h}$ is then transformed by using the inverse function g^r_θ and repeated with this concatenation step $L-1$ times to finally generate the target face image $\mathbf{x}_\mathcal{T}$. An unsqueezed operation is performed to restore the shape of target mouth to match with that of original image sample. Notably, the same flow models in reverse direction g^i_θ and g^r_θ and the same model structure as the training stage are employed in this conditional generation.

E. Disentanglement by the Combined Flow

This article has presented two approaches to flow-based latent disentanglement for conditional generation and mapping from a source image $\mathbf{x}_\mathcal{S}$ (or input image $\mathbf{x}$) to a target or output image $\mathbf{x}_\mathcal{T}$ driven by a query image $\mathbf{x}_q$ (or a condition subimage of lip movement $\mathbf{c}$) where the paired data between source and target domains in face reenactment are missing. The AF flow and the MI flow are proposed by minimizing the structural loss and the information-preserving loss, respectively, toward inferring the attribute-relevant and irrelevant vectors $\{\mathbf{z}^r, \mathbf{z}^i\}$ in flow-based latent representation. Basically, AF flow carries out a two-stage separate training of a flow model f_θ and a disentanglement model with two encoders $\{E_{\phi_r}, E_{\phi_i}\}$ [or a single encoder E_ϕ via Fig. 4(b)] where the likelihood-based loss $\mathcal{L}_f$, structural losses in source and target domains $\{\mathcal{L}_s, \mathcal{L}_t\}$, random-pair loss $\mathcal{L}_p$, and word-level classification loss $\mathcal{L}_y$ are minimized. A flow model is built as a pretrained model, which is then fine-tuned to estimate the encoder. Alternatively, MI flow implements a single-stage disentanglement in the presence of Gaussian prior p_φ of relevant vector $\mathbf{z}^r$ where the informative disentanglement is performed. The variational upper bound of MI of $\{\mathbf{z}^r, \mathbf{z}^i\}$ conditioned on the source mouth $\mathbf{x}$ and his/her lip movement $\mathbf{c}$ is minimized. This bound is factorized as the bounds for domain mapping $\mathcal{L}_d$, condition irrelevance $\mathcal{L}_i$, and conditional relevance $\mathcal{L}_r$.

Basically, two approaches are originated from different perspectives and eligible to be combined to strengthen the flow-based disentanglement with both geometry and information meanings. The combined AF-MI flow is here proposed for latent disentanglement. This hybrid model is implemented by training MI flow as an initial model, which is then fine-tuned in accordance with the objectives of AF flow. A single-stage factorization of $\mathbf{z}^r$ and $\mathbf{z}^i$ is directly handled by a single flow model f_θ (with f^r_θ and f^i_θ) instead of two-stage disentanglement using both flow model f_θ and encoder E_ϕ. The flow parameter θ is updated by jointly minimizing the structural losses in the source domain $\mathcal{L}_s$ and target domain $\mathcal{L}_t$ and the reconstruction loss due to random pairs $\mathcal{L}_p$ in self-supervised manner. Importantly, the word-level classification loss $\mathcal{L}_y$ is minimized. After that, the Gaussian prior p_φ is finally updated by the objective $\mathcal{L}_{\text{MI}}$. In the generation stage, a target image $\mathbf{x}_\mathcal{T}$ is generated from a source mouth $\mathbf{x}_\mathcal{S}$ and his/her lip condition $\mathbf{c}_\mathcal{S}$ by using the flow model and the conditional prior model, which minimizes both $\mathcal{L}_{\text{MI}}$ and $\mathcal{L}_{\text{AF}}$. In this study,

Fig. 7. Some examples in the LRW dataset.

both objectives are incorporated in $\mathcal{L} = \mathcal{L}_{\text{MI}} + \beta\mathcal{L}_{\text{AF}}$ with a hyperparameter β.

V. Experiments

In the experiments, the conditional generation and domain mapping were implemented for mouth reenactment where the flow-based disentanglement is evaluated by using the Oxford-BBC Lip Reading in the Wild (LRW) dataset [52] with some face images shown in Fig. 7.

A. Experimental Setup

The LRW dataset consisted of short human talking videos where each video contained a pronunciation of a single vocabulary word with a length of 29 frames. There were 500 different words that were spoken by hundreds of different speakers. Each word had 1000 utterances. The image size 64×64 with RGB channels was fixed. The audio signals were ignored in this study. The settings for training, validation, and test were referred to [52]. In addition to the proposed AF flow, MI flow, and the combined AF-MI flow, this study also carried out the related works based on the disentangled audio-visual system (DAVS) without and with adversarial learning [15], and the Glow [18] for comparison. AF and MI flow models were implemented with the generative flow based on Glow [18]. Glow was geared with the invertible computation, which estimated the exact likelihood where each flow step implemented three calculations. First, the activation normalization was calculated to act as the scaling layer given by data-dependent initialization. This computation was different from the affine transformation using scale and bias parameters per channel. Second, the invertible 1×1 convolution was calculated, while the dimensions were swapped. This process was different from [16], which simplifies inverted each flow step, and [26], which randomly scrambled the channels. A learnable invertible 1×1 convolution was performed with a generalized permutation by using a rotation matrix with randomly initialized weights [18]. Third, an affine coupling layer [16], [26] was introduced as an invertible transformation where the determinant was computationally efficient. In addition to evaluate the synthesized images, the quantitative evaluation for image reconstruction of the test images in LRW was analyzed in terms of the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) [53], which were averaged over all test images. PSNR measures the ratio of the maximum possible power of color images with RGB channels to the corrupting noise power in decibels. This measure reflects the quality of an image. The higher the better. SSIM measures the similarity between an undistorted image and a distorted image where the factors of luminance, contrast, and structure are jointly considered with equal weighting. This metric is in line with human judgment and is seen as a measure of image quality. The higher the better. The computation time of running Python codes in PyTorch was based on the hardware using a GPU with GeForce RTX 3090 Ti 24 GB and a CPU with Intel Core i9-10900K where a memory of DDR4 128G RAM was used.

Using AF flow, the first stage was to build four computation blocks for multiscale architecture. Each block had 32 flow steps. The mini-batch size was 16. The second stage was to implement two encoders E_{ϕ_i} and E_{ϕ_r} (or a single encoder E_ϕ) and one classifier C_ψ. Parameter ψ was used to transform from the talking attribute vector $\mathbf{z}^r$ with a dimension 6144 ($32 \times 32 \times 6$) to the posterior vector for 500 word labels. Each encoder had one hidden layer. The sequence of frames from a video was used as a batch. The total number of neurons in each layer was the same as the size of image D. Ablation study on the effect of removing individual losses $\mathcal{L}_t$, $\mathcal{L}_p$, and $\mathcal{L}_y$ was evaluated. Using MI flow, there were four computation blocks consisting of three blocks and one block for calculating relevant vector $\mathbf{z}^r$ and irrelevant vector $\mathbf{z}^i$, respectively. MI flow had two components. One was the generative flow f_θ and the other was the conditional prior p_φ. Different from [18], [26], MI flow was implemented by optimizing different MI objectives where a learnable conditional prior distribution $p_\varphi(\mathbf{z}^r|\mathbf{c})$ was merged to establish a multiscale architecture. This conditional prior model served as an encoder to infer the condition-relevant latent variable $\mathbf{z}^r_{\mathcal{S}}$ from a source of subimage of lip movement $\mathbf{c}_{\mathcal{S}}$, which promoted the latent disentanglement for generation of target face $\mathbf{x}_{\mathcal{T}}$ by combining with the condition-irrelevant variable $\mathbf{z}^i_q$ of a query image $\mathbf{x}_q$ based on the trained flow models f^i_θ and f^r_θ. The generation was based on the inverse flow model g_θ. The mini-batch size was set as 8.

In implementation of Glow [18], AF, and MI flows, the dimensions of input and output in the flow model should be the same so as to preserve the invertibility for precise reconstruction. To mitigate the dimensional waste, the multiscale architecture [18], [26] was employed in the flow model by applying the dimensional splitting. In the case of three flow blocks (or $L = 4$), the compressed ratio (denoted by γ) of the dimensions of condition-irrelevant variable $\mathbf{z}^i$ relative to condition-relevant variable $\mathbf{z}^r$ turned out as $\mathbf{z}^i : \mathbf{z}^r = 1 : 2^3$, which resulted in $\gamma = 0.125$. For ablation study, the flow models were implemented with different compressed ratios $\gamma = 0.25$ ($L = 3$) and $\gamma = 0.0625$ ($L = 5$). Nevertheless, MI flow optimized the condition-irrelevant MI $\mathcal{L}_i$, which was able to strengthen the latent variable $\mathbf{z}_i$ with image structural information. Given the condition of mouth movement $\mathbf{c}$, latent variable $\mathbf{z}^r$ was enhanced by optimizing $\mathcal{L}_d$ and $\mathcal{L}_r$ for providing attribute information. In addition, the simplified variant of MI flow was implemented to investigate the effect of different MI terms to infer $\mathbf{z}^i$ and $\mathbf{z}^r$. The ablation study on individually removing $\mathcal{L}_d$, $\mathcal{L}_i$, and $\mathcal{L}_r$ is evaluated. For AF, MI, and AF-MI flows, the bounding parameter $b = 1$ was set. Adam optimizer [54] was used with initial learning rate 0.0001 when updating the parameters θ, ϕ, ψ, and φ. Gradient clipping was applied. The 200k iterations were run.

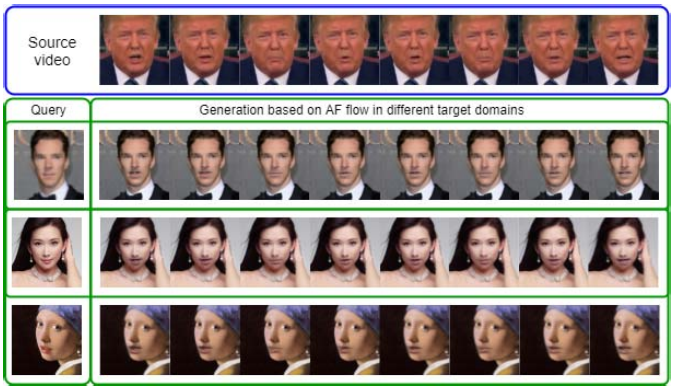

Fig. 8. Face reenactment from a source video and three query images covering different genders, races, and styles. Out-domain data are evaluated. AF flow is applied.

Fig. 9. Face reenactment with ablation study on various loss terms is illustrated. AF flow is applied.

B. Evaluation for Qualitative Results

First, the qualitative evaluation is illustrated for the generated images in the target domain based on the learned flow-based disentanglement for domain mapping from the source images. A source video $\{\mathbf{x}_{\mathcal{S}_n}\}_{n=1}^N$ and a query image $\mathbf{x}_q$ are used to generate the target video $\{\mathbf{x}_{\mathcal{T}_n}\}_{n=1}^N$ frame by frame. A mask is used to crop the region or the subimage around the mouth as the source condition $\mathbf{c}_{\mathcal{S}}$, which is mapped to the corresponding mouth region for query image $\mathbf{x}_q$ to synthesize a target face $\mathbf{x}_{\mathcal{T}}$. It looks like the query image providing the condition of source mouth with his/her lip movement. The scheme of Poisson blending [55] is applied to improve the image quality by tackling the blurred issue. Fig. 8 shows the generated target images $\mathbf{x}_{\mathcal{T}}$ consisting of seven sequence frames where the source images $\mathbf{x}_{\mathcal{S}}$ and a query image $\mathbf{x}_q$ are provided. Latent disentanglement using AF flow is applied. To investigate the generalization capability, the trained model from the LRW dataset is applied for conditional generation of out-domain data where the source videos are collected from YouTube and the query images are sampled from Getty Images (https://www.gettyimages.comhttps://www.gettyimages.com/). Both source and query data are outside the LRW dataset. The evaluation over male/female, western/eastern, and photograph/painting is shown. In general, AF flow obtains desirable imitation for lip movement in target images conditioned on various query faces. The generalization over different genders, races, styles, and angles works well. Fig. 9 further evaluates the mouth reenactment for the other source video where the ablation study on individual loss terms is conducted. Among different losses, it is found that there is no clear change in the synthesized images caused by different source frames when the random-pair reconstruction loss $\mathcal{L}_p$ is removed. In particular, the synthesized mouth in the last two frames does not really reflect the closing mouth as seen in the source video. This implies the importance of $\mathcal{L}_p$ in shaping up the details of lip movement. This loss substantially affects conditional generation and mapping for talking faces.

Next, MI flow is examined for domain mapping. Again, the query images are all excluded from the LRW dataset. The input data consist of a source video $\{\mathbf{x}_{\mathcal{S}_n}\}_{n=1}^N$ or $\{\mathbf{x}_n\}_{n=1}^N$ and a query or condition image $\mathbf{c}$. Fig. 10(a) shows the comparison of the synthesized images for a male and a female. The hyperparameter for tuning MI terms $\alpha = 0.1$ and $\alpha = 0$, where the condition-relevant MI $\mathcal{L}_r$ is functioned and ignored, respectively, is investigated. MI flow with $\mathcal{L}_r$ does provide richer information for better generation. Without $\mathcal{L}_r$, the imitation of lip motion is not obvious. With a smaller value of α, the synthesized lip images are closer to those of source mouth. The condition-relevant variable $\mathbf{z}^r$ does work. Hereafter, $\alpha = 0.1$ is used. In addition, Fig. 10(b) compares the results of the generated videos based on MI flow where different query images with different races or even different portrait paintings are investigated: evaluation for different query faces, including western faces, which are different but close to training targets, and eastern faces and portrait painting faces, which are far from the training targets. As we can see, the sequences of the generated images of lip movement in different races and styles consistently look well. The qualitative results on various out-domain examples assure the generalization performance of AF and MI flows for domain mapping and disentanglement.

C. Evaluation for Quantitative Results

The performance of conditional generation is further evaluated by image reconstruction in terms of PSNR and SSIM that are averaged over the test images of LRW dataset. The baseline results of the DAVS without and with adversarial learning [15] and the Glow model [18] are included for comparison. The DAVS with GAN is examined. The flow model using Glow implements the conditional generation where latent disentanglement is missing. Note that the flow model performs complete reconstruction with the invertible property. The results of latent disentanglement using individual AF and MI flows, and combined AF-MI flow are compared. The ablation studies on various compressed ratios, loss terms, and training styles are investigated. Following the scheme [15] for improving PSNR and SSIM in image reconstruction, the

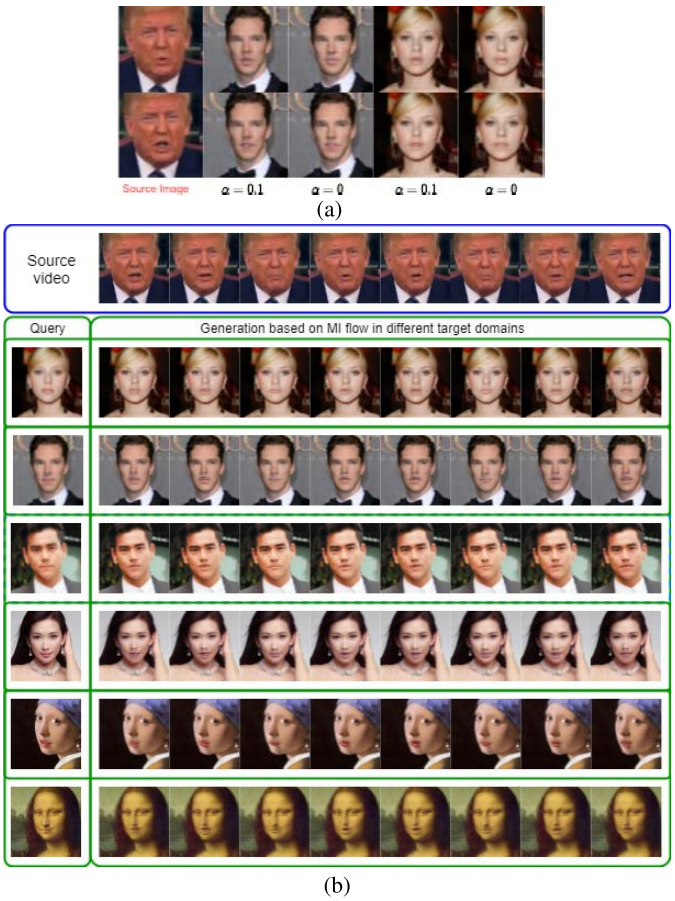

Fig. 10. Face reenactment for (a) evaluation of query images over different genders and hyperparameters α and (b) evaluation of query images covering different races and styles. Out-domain data are evaluated. MI flow is applied.

human face is generated by using a fixed $\mathbf{z}^i$ randomly selected from a video and the $\mathbf{z}^r$ inferred at different frames of the video. Table I compares the PSNR and SSIM scores over different conditional generation models. The training costs of different disentanglement methods relative to Glow under different conditions are reported. In addition to the two-stage implementation of AF flow (denoted as the AF flow-2), AF flow is also implemented by a single-stage AF model (AF flow-1) where the flow model using Glow f_θ and the disentanglement model with encoder E_ϕ are jointly trained instead of treating flow model as a pretrained model for fine-tuning the encoder model. The variants of AF flow in the presence of two encoders $\{E_{\phi_r}, E_{\phi_i}\}$ and a single encoder E_ϕ are also compared. The compressed ratio in AF flow is set as $\gamma = 0.125$. It is found that two-stage AF flow obtains PSNR 28.6 and SSIM 0.939 that are considerably higher than PSNR 25.6 and SSIM 0.918 by using single-stage AF flow. The training procedure via single-stage AF flow does not converge well. In addition, the AF for disentanglement using single encoder E_ϕ performs better than that using individual encoders $\{E_{\phi_r}, E_{\phi_i}\}$ for attribute relevant and irrelevant vectors where PSNR 27.5 and SSIM 0.930 are measured. These variants of AF flow-2 consistently perform better than the baseline systems of DAVS and Glow. Furthermore, the ablation study on learning objectives shows that the performance is dropped

TABLE I

COMPARISON OF PSNR AND SSIM SCORES FOR IMAGE RECONSTRUCTION BY USING DIFFERENT MODELS WITH ABLATION STUDIES ON COMPRESSED RATIO, LOSS FUNCTION, AND TRAINING STYLES. THE TRAINING TIME RELATIVE TO GLOW IS EVALUATED. THE ERROR BAR WITH ONE STANDARD DEVIATION IS SHOWN

Model	PSNR	SSIM	Time
DAVS (w/o advers) [15]	25.7	0.865	–
DAVS [15]	**26.8**	**0.884**	–
Glow ($\gamma = 0.0625$) [18]	**26.0**±0.20	0.892±0.018	1×
Glow ($\gamma = 0.125$) [18]	25.1±0.18	**0.897**±0.019	1.29×
AF flow-1 (E_ϕ)	25.6±0.19	0.918±0.018	1.44×
AF flow-2 (E_ϕ,w/o $\mathcal{L}_s$)	26.3±0.18	0.928±0.019	1.30×
AF flow-2 (E_ϕ,w/o $\mathcal{L}_t$)	26.2±0.22	0.925±0.020	1.32×
AF flow-2 (E_ϕ,w/o $\mathcal{L}_p$)	24.4±0.19	0.886±0.022	1.32×
AF flow-2 (E_ϕ,w/o $\mathcal{L}_y$)	25.4±0.23	0.912±0.019	1.33×
AF flow-2 (E_{ϕ_r},E_{ϕ_i})	27.5±0.20	0.930±0.018	1.43×
AF flow-2 (E_ϕ)	**28.6**±0.19	**0.939**±0.021	1.40×
MI flow ($\mathcal{L}_y$,w/o $\mathcal{L}_d$)	27.0±0.18	0.919±0.019	1.38×
MI flow ($\mathcal{L}_y$,w/o $\mathcal{L}_i$)	26.5±0.17	0.909±0.021	1.39×
MI flow ($\mathcal{L}_y$,w/o $\mathcal{L}_r$)	27.2±0.19	0.921±0.024	1.37×
MI flow ($\gamma = 0.125$)	28.4±0.19	0.936±0.012	1.45×
MI flow ($\mathcal{L}_y$,$\gamma = 0.0625$)	27.9±0.21	0.918±0.019	1.29×
MI flow ($\mathcal{L}_y$,$\gamma = 0.125$)	**29.2**±0.20	**0.949**±0.017	1.48×
MI flow ($\mathcal{L}_y$,$\gamma = 0.25$)	28.1±0.23	0.937±0.020	1.72×
AF-MI flow ($\beta = 0.1$)	29.4±0.19	0.953±0.016	2.20×
AF-MI flow ($\beta = 0.5$)	**30.2**±0.19	0.954±0.023	2.19×
AF-MI flow ($\beta = 0.8$)	29.8±0.21	**0.958**±0.020	2.23×

by individual terms in the loss function $\mathcal{L}_{\text{AF}}$. The biggest drop in model learning was due to the removal of self-supervised learning via random-pair reconstruction loss $\mathcal{L}_p$. The random pairs provide crucial information for reconstruction across two domains. Such a loss conveys sufficient evidence for the encoder to learn AF. Also, the cross-entropy loss $\mathcal{L}_y$ from word label k of a talking mouth is influencing attribute disentanglement in the ablation study. The classification loss $\mathcal{L}_y$ is seen as an additional objective, which is feasible to enrich the inference of attribute-relevant vector $\mathbf{z}^r$. The computation costs due to different training styles of stages and encoders are comparable.

We are accordingly motivated by combining the objective $\mathcal{L}_y$ with the MI objectives in the implementation of MI flow. The effect of adding $\mathcal{L}_y$ is evaluated. The performance of MI flow under different compressed ratios γ is investigated. By additionally merging $\mathcal{L}_y$ in MI flow, PSNR and SSIM are increased from 28.4 and 0.936 to 29.2 and 0.949, respectively, where $\gamma = 0.125$ is fixed. Notably, PSNR and SSIM of MI flow with $\mathcal{L}_y$ are higher than those of AF flow. However, the computation cost of using MI flow is increased as well. The ablation study on individual objectives shows that the largest drops in PSNR and SSIM are caused by the objective $\mathcal{L}_i$, which is known as the most influencing factor in learning objective. The condition-irrelevant MI $\mathcal{L}_i$ is required to capture the mouth structure to improve image reconstruction. Compared with the MIs for domain mapping $\mathcal{L}_d$ and conditional relevance $\mathcal{L}_r$, the condition-irrelevant MI $\mathcal{L}_i$ focuses more on controlling the overall structure and characteristic, which are closer to the performance measures based on PSNR and SSIM. In addition, among different compressed ratios, the value $\gamma = 0.125$ achieves the highest PSNR

and SSIM. In this comparison, even though the compressed ratio is increased to $\gamma = 0.25$, MI flow is not improved in terms of PSNR and SSIM. However, the computation cost is increased significantly by increasing the compressed ratio in the flow model. Basically, PSNR and SSIM using MI flow consistently perform better than those using DAVS without and with adversarial training, Glow with different γ, and AF flow with different training styles. Because of the complementary property in using AF and MI flows, this study presents the flow combination for disentanglement, which is investigated by measuring the results of the combined AF-MI flow with different hyperparameters β. In this comparison, the highest PSNR and SSIM are achieved as 30.2 and 0.958 by using AF-MI flow with $\beta = 0.5$ and $\beta = 0.8$, respectively. Nevertheless, the computation cost of implementing AF-MI flow is increased substantially. The training hours of the best setting using Glow, AF, MI, and AF-MI flows are measured as 20.6, 28.9, 30.5, and 45.1, respectively. Finally, a demo video is provided to illustrate different results of video clips by using AF and MI flows for face reenactment shown in Figs. 8 and 10, respectively.[1] Source codes are commented and posted online in this article.[2]

VI. Conclusion

This article has presented the flow-based latent disentanglement to identify the attribute-relevant and attribute-irrelevant latent variables that were employed for domain mapping and conditional generation. The geometric and informative solutions to disentanglement based on the AF flow and the MI flow were proposed, respectively. AF flow trained the flow model and the feature extractor (or encoder) for attribute relevance and attribute irrelevance based on a two-stage method where the Glow model was estimated by maximizing the generative likelihood and the disentanglement model was inferred by minimizing the structural losses within and between domains. The feature encoder was trained to disentangle the latent vectors according to the structural information of the images. The random-pair reconstruction loss via self-supervised learning and the cross-entropy loss for word classification were additionally minimized without the need of paired data. The proposed loss functions made use of the properties of sequence data and identify the related domain information in different sequences. In addition, this study presented the information-theoretic latent disentanglement for flow-based generative model. A kind of end-to-end training was proposed to carry out the conditional generation for domain mapping in mouth reenactment. The condition-irrelevant and condition-relevant latent variables were learned in accordance with the informative objectives for domain mapping and disentanglement. By introducing the conditional prior, these two latent variables were disentangled and embedded with the specific attribute. AF and MI flows were constructed with the multi-scale architecture where the dimensional waste was handled. The hybrid AF-MI flow combining two flow models was further developed by a cascaded implementation. A series of experiments on qualitative and quantitative evaluation of face reenactment showed the merit of the AF and MI for face generation and reconstruction. The objectives of random-paired reconstruction and condition irrelevance considerably affected the learning procedure. The proposed methods will be further investigated by extending to the other types of flow model and the other kinds of technical data under different domain mapping tasks.

[1]demo video: https://www.youtube.com/watch?v=SuxqpKWr6BQ

[2]source codes: https://github.com/NCTUMLlab/Sheng-Jhe-Huang

References

[1] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 700–708.

[2] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 35–51.

[3] J.-T. Chien, "Deep Bayesian multimedia learning," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4791–4793.

[4] J. T. Chien and C. P. Liao, "Maximum confidence hidden Markov modeling for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 606–616, Apr. 2008.

[5] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent.*, 2014.

[6] E. Mathieu, T. Rainforth, N. Siddharth, and Y. W. Teh, "Disentangling disentanglement in variational autoencoders," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4402–4412.

[7] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[8] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2172–2180.

[9] J.-T. Chien and C.-W. Wang, "Hierarchical and self-attended sequence autoencoder," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Mar. 23, 2021, doi: 10.1109/TPAMI.2021.3068187.

[10] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez, "Invertible conditional GANs for image editing," 2016, *arXiv:1611.06355*.

[11] R. Liu, Y. Liu, X. Gong, X. Wang, and H. Li, "Conditional adversarial generative flow for controllable image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7992–8001.

[12] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[13] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9459–9468.

[14] Y. Song, J. Zhu, D. Li, A. Wang, and H. Qi, "Talking face generation by conditional recurrent adversarial network," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 919–925.

[15] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 9299–9306.

[16] L. Dinh, D. Krueger, and Y. Bengio, "NICE: Non-linear independent components estimation," 2014, *arXiv:1410.8516*.

[17] D. Onken, S. W. Fung, X. Li, and L. Ruthotto, "OT-Flow: Fast and accurate continuous normalizing flows via optimal transport," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 10, 2021, pp. 9223–9232.

[18] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 10215–10224.

[19] H. Sun *et al.*, "DUAL-GLOW: Conditional flow-based generative model for modality transfer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10611–10620.

[20] X. Ma, C. Zhou, X. Li, G. Neubig, and E. Hovy, "FlowSeq: Non-autoregressive conditional sequence generation with generative flow," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 4282–4292.

[21] T.-C. Luo and J.-T. Chien, "Variational dialogue generation with normalizing flows," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7778–7782.

[22] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3617–3621.
[23] X. Ma, X. Kong, S. Zhang, and E. Hovy, "MaCow: Masked convolutional generative flow," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5893–5902.
[24] L. Ardizzone, C. Lüth, J. Kruse, C. Rother, and U. Köthe, "Guided image generation with conditional invertible neural networks," 2019, *arXiv:1907.02392*.
[25] C. Winkler, D. Worrall, E. Hoogeboom, and M. Welling, "Learning likelihoods with conditional normalizing flows," 2019, *arXiv:1912.00042*.
[26] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," 2016, *arXiv:1605.08803*.
[27] J. Ho, X. Chen, A. Srinivas, Y. Duan, and P. Abbeel, "Flow++: Improving flow-based generative models with variational dequantization and architecture design," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2722–2730.
[28] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu, "Lip movements generation at a glance," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 520–535.
[29] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou, "Large scale 3D morphable models," *Int. J. Comput. Vis.*, vol. 126, pp. 233–254, Apr. 2018.
[30] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2387–2395.
[31] O. Wiles, A. S. Koepke, and A. Zisserman, "X2Face: A network for controlling face generation using images, audio, and pose codes," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 670–686.
[32] J.-C. Tsai and J.-T. Chien, "Adversarial domain separation and adaptation," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2017, pp. 1–6.
[33] Y. Tu, M.-W. Mak, and J.-T. Chien, "Variational domain adversarial learning with mutual information maximization for speaker verification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2013–2024, 2020.
[34] L. Li, M.-W. Mak, and J.-T. Chien, "Contrastive adversarial domain adaptation networks for speaker recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 5, pp. 2236–2245, May 2022.
[35] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4176–4186.
[36] H.-J. Chen, K.-M. Hui, S.-Y. Wang, L.-W. Tsao, H.-H. Shuai, and W.-H. Cheng, "BeautyGlow: On-demand makeup transfer framework with reversible generative network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10042–10050.
[37] G. Desjardins, A. Courville, and Y. Bengio, "Disentangling factors of variation via generative entangling," 2012, *arXiv:1210.5474*.
[38] I. Higgins *et al.*, "Towards a definition of disentangled representations," 2018, *arXiv:1812.02230*.
[39] R. T. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 2610–2620.
[40] B. Esmaeili *et al.*, "Structured disentangled representations," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2019, pp. 2525–2534.
[41] E. Denton and V. Birodkar, "Unsupervised learning of disentangled representations from video," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4414–4423.
[42] J.-T. Chien and Y.-Y. Lyu, "Partially adversarial learning and adaptation," in *Proc. 27th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2019, pp. 1–5.
[43] H. Kim and A. Mnih, "Disentangling by factorising," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2649–2658.
[44] A. Kumar, P. Sattigeri, and A. Balakrishnan, "Variational inference of disentangled latent concepts from unlabeled observations," in *Proc. Int. Conf. Learn. Represent.*, 2018.
[45] J. Yang, S. E. Reed, M.-H. Yang, and H. Lee, "Weakly-supervised disentangling with recurrent transformations for 3D view synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1099–1107.
[46] P. Sorrenson, C. Rother, and U. Köthe, "Disentanglement by nonlinear ICA with general incompressible-flow networks (GIN)," in *Proc. Int. Conf. Learn. Represent.*, 2020.
[47] R. Kondo, K. Kawano, S. Koide, and T. Kutsuna, "Flow-based image-to-image translation with feature disentanglement," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 4168–4178.
[48] S.-J. Huang and J.-T. Chien, "Attribute decomposition for flow-based domain mapping," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 1710–1714.
[49] F. Zhu *et al.*, "Image-text dual neural network with decision strategy for small-sample image classification," *Neurocomputing*, vol. 328, pp. 182–188, Feb. 2018.
[50] L. Ardizzone, R. Mackowiak, C. Rother, and U. Köthe, "Training normalizing flows with the information bottleneck for competitive generative classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 7828–7840.
[51] I. Jeon, W. Lee, M. Pyeon, and G. Kim, "IB-GAN: Disentangled representation learning with information bottleneck generative adversarial networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 9, 2021, pp. 7926–7934.
[52] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 87–103.
[53] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[55] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *Proc. ACM SIGGRAPH Papers (SIGGRAPH)*, 2003, pp. 313–318.

Jen-Tzung Chien (Senior Member, IEEE) is currently the Lifetime Chair Professor in electrical and computer engineering and computer science at National Yang Ming Chiao Tung University, Hsinchu, Taiwan. He has published extensively, including three books and over 250 peer-reviewed articles, many on machine learning, deep learning, and Bayesian learning with applications on natural language processing and computer vision.

Dr. Chien served as a Tutorial Speaker for AAAI, IJCAI, ACL, KDD, MM, ICASSP, ICME, CIKM, IJCNN, COLING, and Interspeech.

Sheng-Jhe Huang received the B.S. degree in electrical and computer engineering from National Central University, Taoyuan, Taiwan, in 2018, and the M.S. degree in electrical and computer engineering from National Yang Ming Chiao Tung University, Hsinchu, Taiwan, in 2020.

His research interests include domain mapping, image translation, generative model, and information-theoretic learning.