Particularity beyond Commonality: Unpaired Identity Transfer with Multiple References

Ruizheng Wu[1], Xin Tao[2], Yingcong Chen[3], Xiaoyong Shen[4], Jiaya Jia[1,4]
[1]The Chinese University of Hong Kong, [2]Kuaishou Technology,
[3]CSAIL, MIT, [4]SmartMore
{rzwu, leojia}@cse.cuhk.edu.hk, jiangsutx@gmail.com,
ycchen@csail.mit.edu, xiaoyong@smartmore.com

Abstract. Unpaired image-to-image translation aims to translate images from the source class to target one by providing sufficient data for these classes. Current few-shot translation methods use multiple reference images to describe the target domain through extracting common features. In this paper, we focus on a more specific identity transfer problem and advocate that particular property in each individual image can also benefit generation. We accordingly propose a new multi-reference identity transfer framework by simultaneously making use of particularity and commonality of reference. It is achieved via a semantic pyramid alignment module to make proper use of geometric information for individual images, as well as an attention module to aggregate for the final transformation. Extensive experiments demonstrate the effectiveness of our framework given the promising results in a number of identity transfer applications.

1 Introduction

Reference images are usually used as a supplement in image translation and image editing tasks [7, 11, 25, 57]. They provide guidance of content (e.g., poses, expression) [7, 57], style (such as texture) [11] or category information (e.g., identities, expression labels) [3, 25] for the final results. In this paper, we focus on identity transfer tasks, including clothes and face identity transfer.

In certain tasks, multiple reference images as guidance are also available. Image translation frameworks of [25, 51] utilize these multiple inputs, usually unseen classes for generation, proved to be effective to achieve promising results. We term them as *few-shot-based* methods. There is still an enormous room to explore the appropriate way to use multiple references.

Few-Shot *vs.* Multi-Reference Multiple reference images contain variation in many dimensions while keeping one common attribute. For example, in face identity transfer, reference images vary from poses and expression while maintaining the same identity. Few-shot-based methods [25, 51], contrarily, diminish variation in the unconcerned attributes and only focus on *commonality*.

This paper forms a new point of view that *particularity* inside each reference also provides useful clues for generation. Intuitively, in the task of identity transfer, if the reference image shares the similar poses/expression as the input content image, then the desired output can copy a lot of patches from the reference. In [12, 36, 46], the outputs are generated by warping from one reference image in image space. However, these methods are mostly applicable for images with similar poses (e.g. frontal faces), when the pose of a reference image differs greatly from the content image, the warping technique in these methods is very likely to fail. To obtain more robust results, we consider using multiple reference images, as more references can provide complementary for generation. For example, in Fig. 4(b), the 3rd reference image provides a frontal face, and the 2nd reference provides a contemptuous mouth for the final output.

With this new motivation, we propose an intriguing way of using multiple reference images and name it *multi-reference* method. In this method, we obtain *particularity* from individual reference with an alignment module, and we adaptively assign weights with an attention module to references for fusion as *commonality*. The effectiveness of this new line of approach can be well proved on the unpaired identity transfer task.

Semantic Alignment To make full use of *particularity* inside each reference, we carefully design the alignment module for individual references. There exist prior image translation tasks involving alignment, which estimate pixel-wise correspondence in image space [12, 44, 46]. It is widely known that pixel-level alignment among images in different domains may result in unwanted distortion given domains of, for instance, cartoon and real faces. In this case, we believe semantic-level alignment is more important than the pixel-level one, which takes context information into consideration for deeper image understanding.

We thus introduce a semantic pyramid to represent different levels of image features, and a new module named semantic pyramid alignment to align images hierarchically. The module starts from the highest semantic level and progressively refine estimated correspondence in lower levels. Unlike previous multi-level feature matching [1, 23] that mainly searches sparse or dense correspondence in a feature extraction network (e.g., pre-trained VGG), we instead accomplish results by semantic alignment in an end-to-end fashion. Our alignment module empirically outperforms several single-level alignment baselines.

Our contribution is as follows. 1) We propose a multi-reference framework, which takes advantage of priors in each individual reference image, and adaptively fuses and generates result images. 2) We propose a semantic pyramid alignment module, which aligns references semantically in multiple levels. 3) We design an attention module to adaptively assign weights for reference fusion, along with an effective category classification and comparison discriminator to enforce image generation in a specific domain. 4) We achieve promising results on several unpaired identity transfer tasks with only a few reference images.

2 Related Work

Unpaired Image Translation Unpaired Image Translation is a task to translate source domain images to target domain ones without ground-truth. Methods of [18, 50, 57] use cycle consistency loss for image reconstruction, which greatly improves performance without paired data. To translate multiple target domains in one trained network, methods of [7, 33, 46] introduce class conditions as extra input for translation. In another stream, some works [6, 24, 26, 30, 40, 45] address this task with the assumption that images in source and target domains share the same latent space. Moreover, methods of [17, 21, 47] disentangle the latent space into style, structure, and content partition, so that the generator can fuse features from different spaces.

To better utilize the feature from input images, some approaches [8, 12, 37, 46, 52] propose to warp features [8, 37] or pixels [12, 46] from inputs to the translated results. Different from them, our alignment is accomplished on semantic pyramid for higher robustness and it is done on multiple references simultaneously. Similar to our approach, which fuses priors from multiple references, a view synthesis network [56] is also proposed to synthesize a novel view of a scene or an object, while we focus more on translation among different identities.

Recently, research [4, 25, 42, 51] also involves few-shot unpaired image translation, which is more challenging since only one or several target/source domain images are provided for training or inference. Different from the most related work [25], our framework takes both the geometric priors from each reference and global context of all references into consideration, and thus achieves decent results. The method of [42] also tackles temporal alignment and video frame fusion, while we consider more on semantic alignment among different domains and adaptively fuse the clues from references.

Face Image Generation To generate faces with different poses or expressions, research works [2, 10, 12, 16, 36, 44, 46] have been proposed to synthesize face or head images by warping a single or multiple reference images. The work of [2, 12] warp images with face landmark features, while in [10, 36, 44, 46], the input images are warped with learned warping field. But there is a chance to produce distortion or visual artifacts when head rotation, large motion or occlusion exists.

Alternatively, deep convolutional networks were considered [3, 9, 22, 31, 33, 34, 41, 48, 51, 54, 55]. In [9, 33, 34, 41, 48, 51, 54], faces are generated with the guidance of segmentation map, facial landmark, boundary map, or pose/expression parameters. Bao et al. [3] disentangle identity and facial attribute for face synthesis. Natsume et al. [31] swap faces between two identities, while Zakharov et al. [51] synthesize talking heads by adopting few-shot adversarial learning strategy. They do not take full advantage of the geometric clues from multiple input faces. As for the work of Sungjoo et al. [14], they also consider multiple reference images for final generation. They apply adaptive weights for under-aligned feature blending and average the independently aligned feature, while we combine the two steps of aligning different references and adaptive aggregation without the help of extra landmark input.

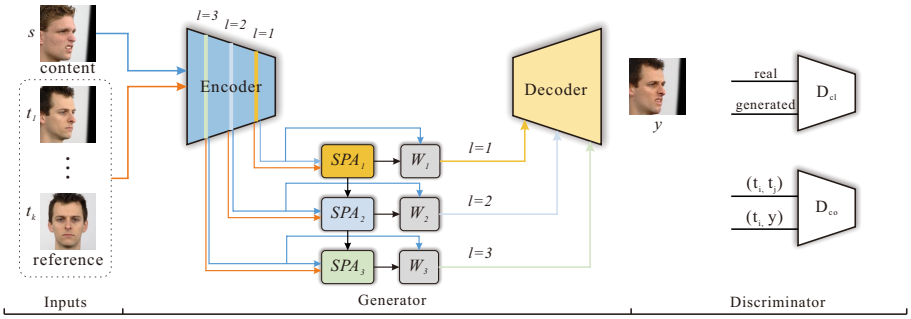


Fig. 1. The overall architecture of our multi-reference framework, where $\{SPA_i\}$ denotes the semantic pyramid alignment module, $\{W_i\}$ indicates the attention module for fusion. l refers to the feature level. All reference images are aligned with the content image in multi-level feature space, and references are adaptively fused with the attention module for the decoder to generate the final result.

3 Our Method

Our method is to translate an image from the source class to the target one under the condition that the target class is unseen in the training set and is only specified by one or a few reference images. Specifically, given one content image s from the source class $\mathbb{S}$ and k reference images $\{t_i\}_{i=1,2,...,k}$ from the target class $\mathbb{T}$, we generate output y in the target class $\mathbb{T}$ while preserving the content of pose, expressions, and shape from s.

To this end, we propose a multi-reference framework (Fig. 1), which makes appropriate use of the clues from references to generate the final result. In the generation stage, we firstly align each reference and the content image by applying the semantic pyramid alignment module, and then adaptively fuse all reference features with an attention sub-network. Finally, we decode the fused features to obtain the output hierarchically [35]. In the discrimination stage, both category classification and comparison are adopted in discriminator establishment.

3.1 Multi-Reference Guided Generator

To utilize multiple reference images, current few-shot image generation frameworks [25, 51] extract a spatially invariant embedding vector for each reference. The vectors are averaged as global context for decoding. Although these methods extract *commonality* from references, *particularity* in each individual reference is discarded, which however also provides vitally important clues for generation.

To address this issue, we consider *particularity* by aligning each individual reference with the content image for generation, without sacrificing *commonality* due to the important weighted fusion. To adaptively obtain global context while retaining clues from each reference, the most important two modules for alignment and fusion are as follows.

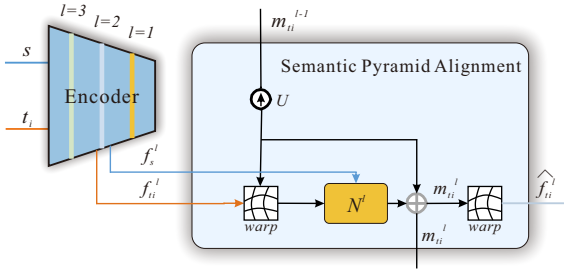


Fig. 2. Median-level structure in semantic pyramid alignment module.

In the alignment module, for each reference, we align it with the content image semantically. Rather than directly aligning on image space, we start from the high-level feature space to learn a coarse semantic correspondence, which is then propagated to lower feature space for refinement. It not only enables the semantic context aligned in high-level feature space, but also preserves textures from references in low-level one. In the fusion stage, we estimate the weight map for reference by attention module in each feature space. Noted that we do not map the image into a 1D embedding vector and instead produce a 2D feature map to preserve the spatial structure of each reference.

Semantic Pyramid Alignment As mentioned above, we align the reference and content images in multi-level deep feature space, which is referred to as semantic pyramid. The pipeline is described in Alg. 1, where l refers to different levels, and $l = 1, 2, 3$ indicates high-, median- and low-level feature space. Besides, i indexes reference images with number k in total.

Starting from the highest level, we estimate feature correspondence between current and each reference to obtain the coarsest optical flow map, which is then fed into the lower level for refinement. For subsequent median and low levels, we update the coarse optical flow by estimating a residual.

The procedure is to first up-sample the optical flow map $\{m_{ti}^{l-1}\}$ from the high level. It is used to warp the current reference feature f_{ti}^l to $w(f_{ti}^l)$ that is coarsely aligned with content f_s^l. We then estimate the dense correspondence between $w(f_{ti}^l)$ and f_s^l in a finer level with the alignment network N^l. The network output is a residual for the upsampled flow map from the last level. We sum them up to get the final flow map $\widehat{f_{ti}^l}$. The structure of median-level alignment module is shown in Fig. 2. With optical flow maps estimated on multiple levels, the reference image feature is warped for our deployment.

Feature Fusion with Attention By aligning features from all reference images, we select useful regions for generation. Since output image preserves pose, expressions, and shape from the content image, we thus search similar patches with the content image from all references.

Algorithm 1: Semantic Pyramid Alignment

Input: Alignment network N^l, content image feature f_s^l, reference image feature f_{ti}^l, where $l = 1, 2, 3$ and $i = 1, 2, ..., k$.
Output: Flow maps $\{m_{ti}^l\}$, warped features $\{\widehat{f_{ti}^l}\}$.

```
for l = 1; l ≤ 3; l = l + 1 do
    if l == 1 then
        for i = 1; i ≤ k; i = i + 1 do
            m_ti^l = N^l(f_s^l, f_ti^l);
        end
    else
        for i = 1; i ≤ k; i = i + 1 do
            U(m_ti^{l-1}) = Upsample(m_ti^{l-1}) × 2.0 ;
            w(f_ti^l) = warp(f_ti^l, U(m_ti^{l-1})) ;
            m_ti^l = N^l(f_s^l, w(f_ti^l)) + U(m_ti^{l-1});
        end
    end
    f̂_ti^l = warp(f_ti^l, m_ti^l) ;
end
return {m_ti^l}, {f̂_ti^l};
```

For each level in feature space, as shown in Fig. 3, we firstly flatten and transpose the content feature f_s^l and all reference ones $\widehat{f_t^l}$ into an appropriate shape, then we calculate their similarity at each pixel location by batch matrix multiplication. This operation results in attention weights for each reference, and then they are normalized with *softmax* on the dimension of reference number. Finally, we apply the normalized weights on reference features $\widehat{f_t^l}$ and obtain the output o^l by another matrix multiplication for fusion. o^l is fed to decoder for final generation. Noted that it is related but different from self-attention blocks [43, 53, 49]. In self-attention blocks, the similarity matrix is computed among features of pixels, while we compute the similarity among features of different references.

3.2 Discriminators

To distinguish the generation of the translated images from the real ones, we employ two discriminators, i.e. category classification discriminator D_{cl} and category comparison discriminator D_{co}.

Category Classification Following the discriminator in [25], we build a multi-task adversarial discriminator to distinguish between the generated and real images with multiple categories. For $|\mathbb{S}|$ categories in the training set, the discriminator D_{cl} produces $|\mathbb{S}|$ output, and we treat each as binary classification.

The adversarial loss is applied to specific class output. When updating D_{cl} for a real image of source class s, we penalize D_{cl} if its sth output is negative.

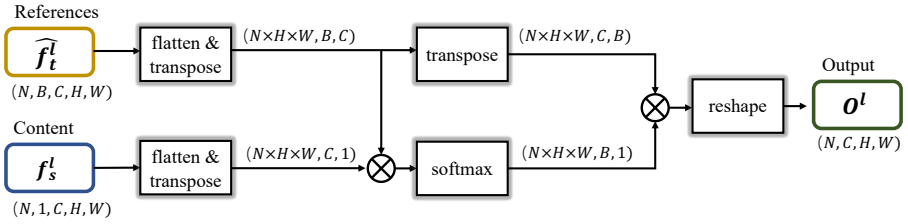


Fig. 3. The procedure of feature fusion in attention mechanism, where n, b, c, h and w refer to batch size, reference number, channel, height, and width respectively.

For fake images of class s, we penalize D_{cl} if its sth output is positive. As for generator, we only penalize generator if the sth output of D_{cl} is negative.

Category Comparison Only category classification discriminator is not enough to enforce generation of unseen categories, since the classification discriminator is trained only with the known categories in the training set. When an image from an unseen category is required to be generated, the generator needs a similar category in the training dataset. Therefore, in the case that there is no similar category, we need an auxiliary network to strengthen the relation between reference inputs and output.

To this end, we design another discriminator D_{co} for comparing the category between two images. We treat two images as a positive sample if they belong to the same category. Otherwise, they are negative. Besides, the generated and real images in the same category are viewed as another negative sample. The discriminator helps preserve the identity in unseen classes.

3.3 Training

Random Warping for Reference Images Directly training of our proposed framework can be easily trapped into generating one of the most similar reference images. To avoid this trivial solution, which copies one of the reference images, we randomly produce the warping parameters and apply them to distort the reference images as a perturbation.

Loss Functions We adopt two adversarial losses $\mathcal{L}_{GAN}^{cl}$ and $\mathcal{L}_{GAN}^{co}$ for two kinds of discriminators of D_{cl} and D_{co}. D_{cl} is a $|\mathbb{S}|$ binary discriminator, and D_{co} is a conditional binary discriminator. The two loss functions are

$$\mathcal{L}_{GAN}^{cl}(G, D_{cl}) = \mathbb{E}_{\mathbf{s}}[-logD_{cl}(\mathbf{s})] + \mathbb{E}_{\mathbf{s},\{\mathbf{t_1},\ldots,\mathbf{t_k}\}}[log(1 - D_{cl}(\mathbf{y}))],$$
$$\mathcal{L}_{GAN}^{co}(G, D_{co}) = \mathbb{E}_{\mathbf{t_i},\mathbf{t_j}}[-logD_{co}(\mathbf{t_i}, \mathbf{t_j})] + \mathbb{E}_{\mathbf{s},\{\mathbf{t_1},\ldots,\mathbf{t_k}\}}[log(1 - D_{cl}(G(\mathbf{t_i}, \mathbf{y})))],$$

Where s, $\{t_i\}$ and y indicate source class image, target image set, and translated image respectively.

We also adopt reconstruction loss by sampling content and reference images in the same category. In this case, the output image is the same as the content one. It is expressed as

$$\mathcal{L}_{REC}(G) = \mathbb{E}_{\mathbf{s},\{\mathbf{s_1},...,\mathbf{s_k}\}}\|\mathbf{s} - G(\mathbf{s}, \{\mathbf{s_1}, ..., \mathbf{s_k}\})\|_1, \tag{1}$$

where $\{\mathbf{s_1}, ..., \mathbf{s_k}\}$ indicates a set of k random samples from category $\mathbb{S}$.

4 Experiments

4.1 Datasets and Implementation

Datasets To verify the effectiveness of our framework, we conduct experiments on two kinds of datasets of faces and human body. Face/clothes identity transfer tasks are accomplished. For face identity transfer, we conduct experiments on RaFD [20], Multi-PIE [13] and CelebA [29], while we utilize DeepFashion [28] for clothes identity transfer.

Both RaFD and Multi-PIE contain face images with a clean background. RaFD contains 67 identities, and we use the first session of Multi-PIE with 249 identities. CelebA is a more challenging dataset with complicated background, which contains 10,177 identities, while different images of the same person may vary widely. As for Deepfashion dataset, we utilize its group of 'Blouses_Shirts', and we split them into 1,438 styles for training and 189 for test.

Implementation We implement our method with PyTorch [32] on a TITAN Xp card. We train our framework with resolution 128×128, and set batch size as 6. Adam [19] optimizer with learning rate $1e-4$ is adopted for both generator and discriminators. We train our framework with 3 reference images (3-shot). Any number of references can be fed into our framework for testing. At the stage of inference, the class/identity of reference images is unseen in training.

4.2 Quantitative Evaluation Metrics

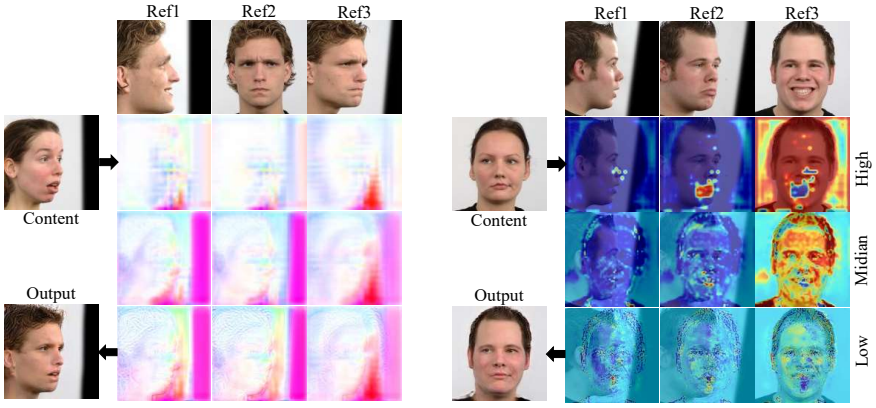
We set up quantitative evaluation metrics as follows.

Classification Accuracy (Acc) Similar to that of [7, 25], we train a classifier for testset. We adopt a pre-trained Inception-v3 [39] as backbone and replace the fully connected layer with a new one that produce specific class number for corresponding dataset. We evaluate classification accuracy of generated images.

Distribution Discrepancy (mFID) To obtain distribution discrepancy, we firstly extract features with a deep face feature extractor VGGFace2 [5] for face dataset and VGG16 [38] for DeepFashion dataset, then we use FID [15] to measure feature distribution discrepancy between real and generated faces for each category and obtain the average as mFID.

Inception Score (IS) We utilize the fine-tuned Inception-v3 [39] in 'Classification Accuracy' to calculate the inception score between real and generated images, which measures the realism of generated images.

Perceptual Distance (Per) We also measure if our results preserve the content by calculating the $\mathcal{L}_2$ distance between the content image and output in feature space produced by the 'conv_5' layer in pre-trained VGG16 on ImageNet.



(a) Optical flow visualization among different level features, 'Ref*i*' indicates the *i*th reference. For each reference, we visualize the corresponding optical flow from high to low level feature space in 3 rows.

(b) Attention weight visualization (blue → low, red→ high) on different levels. Similar to (a), we show the visualization of images and low-, median- and high-level attention maps.

Fig. 4. Visualization of (a) pyramid alignment and (b) attention maps for fusion.

4.3 Analysis of Different Components

Visualization of Pyramid Alignment Module In our proposed framework, hierarchical features from the encoder are aligned by estimating correspondence between content and reference images. The dense correspondence on multiple levels is visualized in Fig. 4(a). We notice that optical flow (dense correspondence) is gradually refined from high to low levels.

Visualization of Attention Map for Fusion The attention weight on each level for selection and fusion is visualized in Fig. 4(b). The attention module produces different weights on regions – regions that are similar between the reference and the content images gain large attention weights. For example, in the high level space, most face regions of 'Ref3' receive great weight as they share similar pose with the content image; the mouth of 'Ref2' is also important with the same reason.

Besides, from high to low levels, attention weights dilute for each reference image, since features are gradually aligned and constructed by our alignment module and decoder layers.

Semantic Pyramid *vs.* Single Level In the alignment and attention module, we utilize multiple level features in a hierarchical mechanism. To verify the effectiveness, we compare with simplified versions with only one level. The versions include the highest level feature (i.e. the last layer of encoder) for alignment and fusion and the lowest level where alignment and fusion are directly done on RGB image space. The results shown in Fig. 5 indicate that the output is easily trapped due to model collapse with only the highest level feature, while distortion emerges when processing on RGB image level.

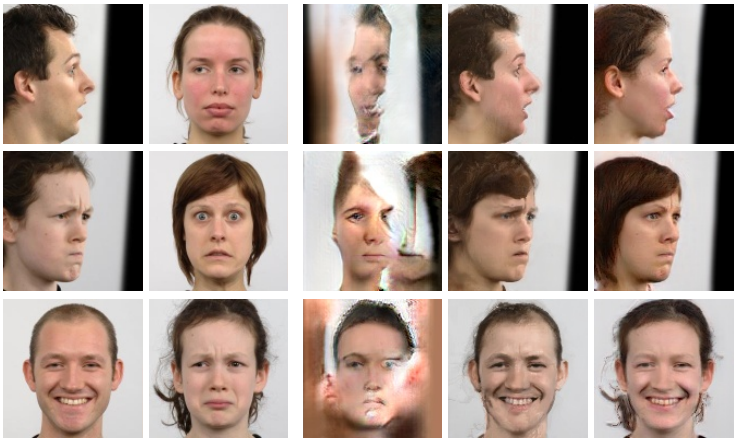


Fig. 5. Results of multi-level and simplified single-level frameworks, where 'Single_H' and 'Single_R' means that alignment is conducted on only the highest feature and RGB-image levels respectively. 'Multi' indicates our complete semantic pyramid alignment.

Effectiveness of Discriminators We utilize two discriminators for category classification and comparison. To study the roles of each discriminator, we conduct ablation study on discriminators. The quantitative results are shown in Table 1, where 'w/o D_{co}', 'w/o D_{cl}' and 'Full' indicate the proposed framework without category classification discriminator, framework without category comparison discriminator and our full system.

In Table 1, we observe that without the category classification module, translation accuracy drops significantly while the inception score gets better. It means that category comparison is beneficial for higher-quality image generation, while category classification helps specific category image generation. Note that the translation accuracy and mean FID reach the best values when we utilize both.

Reference Quality and Reference Number Since we make use of clues from references for generation, the quality of references makes difference for final generation. We conduct experiments on RaFD dataset, and sample images with different poses and expressions as reference images. Results in Fig. 6 show that FUNIT [25] differs little even with the change of references, while our proposed

Table 1. Quantitative comparison among different discriminators on RaFD dataset.

	Acc(%) ↑	IS ↑	mFID(1e3) ↓	Per ↓
w/o D_{co}	68.84	2.41	7.70	1.41
w/o D_{cl}	16.77	**5.80**	12.13	**0.94**
Full	**73.03**	2.49	**6.08**	1.39

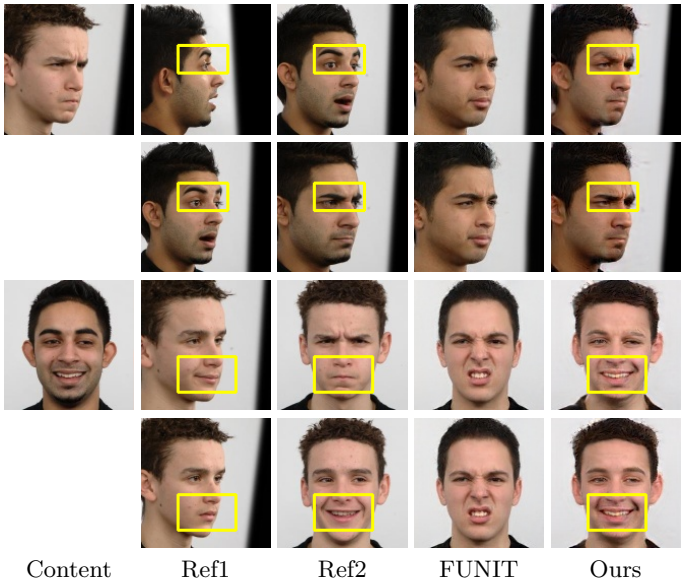


Fig. 6. Effect of reference quality for final generation. The highlight regions show the main effect with different references. In the first case, similar eyebrow between 'Ref2' and content leads to better eyebrow in translation (2nd row). In the second example, smile expression in 'Ref2' helps the generation of smiling face (4th row).

framework generates better results with more suitable reference images. The more similar poses content and reference images are, the higher-quality images we translate. This attributes to the fact that similar poses provide realistic clues for final generation.

We then evaluate our framework with different reference numbers. In the experiments, we randomly select reference images to specific number. They are fed into a trained framework. In Table 2, the inception score (IS) and perceptual distance (Per) are comparable, while scores of classification accuracy (Acc) and mFID improve greatly with reference number increasing, indicating that more references help achieve more accurate translation.

Table 2. Quantitative comparison regarding reference numbers.

Ref	Acc(%) ↑	IS ↑	mFID(1e3) ↓	Per ↓	Runtime(ms) ↓
1	53.12	**3.31**	9.55	1.56	**51.1**
3	73.03	2.49	6.08	**1.39**	53.0
5	90.01	1.46	4.50	1.47	59.8
10	**90.21**	1.47	**4.39**	1.44	71.5

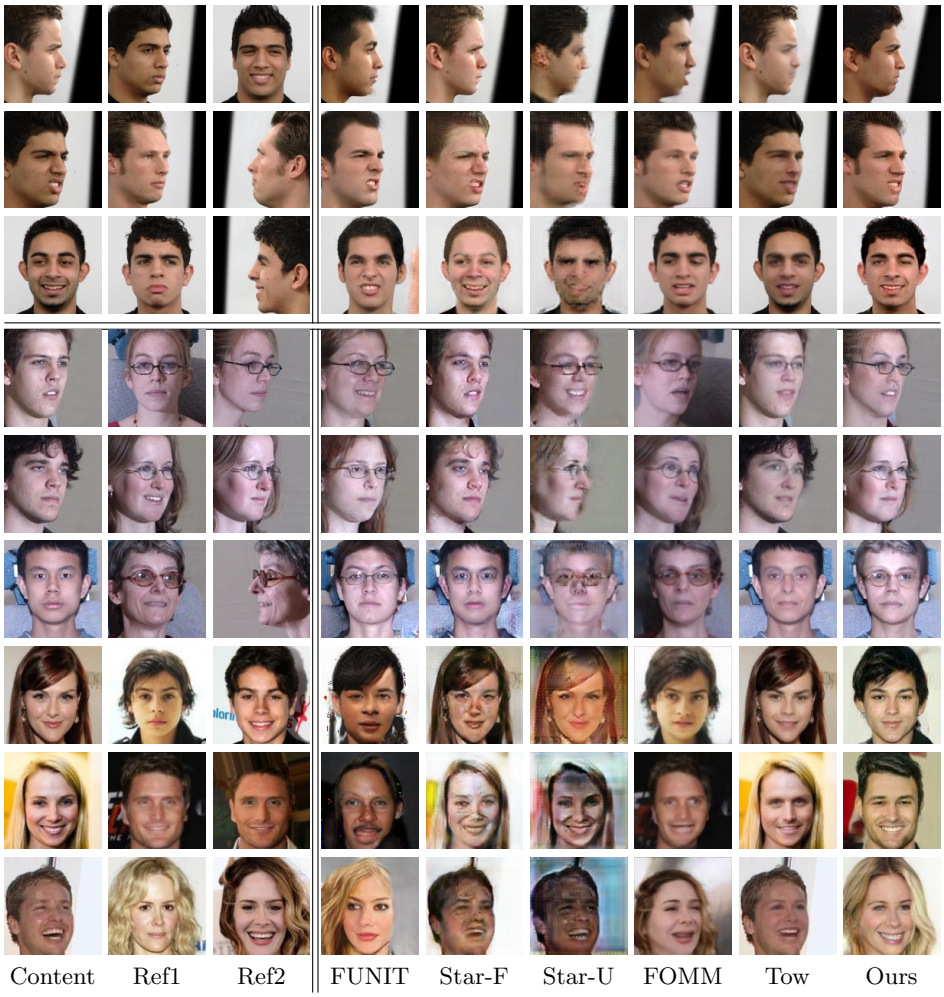


Fig. 7. Visual comparison on different face datasets. The results are generated with 3 references, while only 2 references are shown here.

4.4 More Results

Identity Transfer on Faces For face identity transfer, we conduct experiments on RaFD [20], Multi-PIE [13] and CelebA [29]. We compare with approaches of FUNIT [25], state-of-the-art few-shot unpaired image translation framework; Tow [3], the method to synthesize faces with identity and attributes from two faces respectively, and FOMM [37], the latest motion transfer method. We also compare with StarGAN [7], state-of-the-art multi-class unpaired image translation method. Since StarGAN is a full-shot framework, we follow the 'fair' ('Star-F') and 'unfair' ('Star-U') setting of [25].

Fig. 8. Visual comparison on clothes identity transfer. We resize PATN results from 256×176 to 128×128, which make them slightly misaligned with the input.

As for Star-F, we firstly train StarGAN with seen categories in the training dataset. During testing, we estimate class association vectors [25] for input reference images. Then the estimated vectors are used as the target condition for inference. For the 'unfair' setting of StarGAN ('Star-U'), we train the framework with images of unseen categories in the test set. We adopt 3-shot training in our experiments, i.e., 3 reference images are provided in training.

Visual Comparison We show the visual comparison in Fig. 7. As for the results on RaFD and Multi-PIE dataset, Star-F produces results with correct pose and expression. Yet the identity is incorrect. Contrarily, Star-U achieves results with correct identities; but the output is blurry with more visual artifacts. They are mainly caused by a lack of sufficient training data with only 3 reference samples provided. FOMM generates satisified results when reference and content image share similar poses, while it generates distorted results in the cases that large geometric transform is required. As for Tow, center faces are translated to another identity but the results are kind of blurry. FUNIT generates decent output with fewer artifacts, and the identity is similar to one person in the training set, while it fails to generate images with the satisfying categories specified by the reference images. Since our method makes appropriate use of the clues from each reference, our results preserve identities from references best. Besides, the alignment module in semantic pyramid greatly improves the generation quality.

For CelebA dataset, reference images vary much though they belong to the same identity, and thus it would be difficult to define target identity with only these few references. As a result, Star-F and Star-U cannot generate decent output since there are too many identities but few samples provided for each in the training set. As for FUNIT, it can generate decent output, while our method can produce more promising results with few varied references.

Table 3. Quantitative comparison on RaFD and DeepFashion respectively.

Application	Methods	Acc(%) ↑	IS ↑	mFID ↓	Per ↓
Faces	FUNIT [25]	38.31	1.83	12.40	1.62
	Star-F [7]	16.91	1.82	15.56	1.41
	Star-UN [7]	31.71	1.41	15.34	1.66
	FOMM [37]	33.94	1.39	9.73	1.76
	Ours	**73.03**	**2.49**	**6.08**	**1.39**
Clothes	FUNIT [25]	0.69	10.74	25.96	0.14
	LiquidGAN [27]	**36.9**	2.04	35.19	1.70
	PATN [58]	12.29	2.19	32.24	0.16
	Ours	29.97	**13.38**	**25.45**	**0.12**

Quantitative Comparison We also make comparison quantitatively in Table 3. Our method achieves both the highest classification accuracy and IS score among all compared methods on face dataset. It indicates that our method consistently guarantees correct category and ensures high-quality results. Besides, we achieve the lowest mFID score and perceptual distance with content images, showing that our results also preserve the content well from the input.

Identity Transfer on Clothes For the application of clothes translation on human body, we utilize DeepFashion dataset for method comparison. We evaluate different frameworks regarding the group of 'Blouses_Shirts'. Different styles of clothes are regarded as the categories in our framework.

We compare our method with LiquidGAN [27] and PATN [58], which are used for pose translation. Their pre-trained models on DeepFashion dataset are adopted for test. We also compare with FUNIT here. The results are shown in Fig. 8. Both LiquidGAN and our method translate the input images to the correct categories of clothes, while our generated images are more natural.

The quantitative comparison is listed in Table 3. Since LiquidGAN [27] and PATN [58] are single reference frameworks, we generate results with all 3 testing references and average their evaluation metrics scores. Noted that we can achieve the best scores on both IS and mFID, which indicates the high quality of our results. Besides, we also preserve poses from content images well with the shortest perceptual distance, while LiquidGAN and our method achieve comparable classification accuracy.

5 Conclusion

In this paper, we propose a multi-reference framework for unpaired identity transfer, which makes decent use of clues from each individual reference. A well-designed semantic pyramid alignment module is introduced to extract *particularity* from each reference. References are also adaptively fused as *commonality* for generation with the attention module. We conduct extensive experiments and achieve promising results on some unpaired identity transfer applications.

References

1. Aberman, K., Liao, J., Shi, M., Lischinski, D., Chen, B., Cohen-Or, D.: Neural best-buddies: Sparse cross-domain correspondence. ACM Trans. Graph. (2018)
2. Averbuch-Elor, H., Cohen-Or, D., Kopf, J., Cohen, M.F.: Bringing portraits to life. ACM Trans. Graph. (2017)
3. Bao, J., Chen, D., Wen, F., Li, H., Hua, G.: Towards open-set identity preserving face synthesis. In: CVPR (2018)
4. Benaim, S., Wolf, L.: One-shot unsupervised cross domain translation. In: NeurIPS (2018)
5. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (2018)
6. Chen, Y.C., Xu, X., Tian, Z., Jia, J.: Homomorphic latent space interpolation for unpaired image-to-image translation. In: CVPR (2019)
7. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR (2018)
8. Dong, H., Liang, X., Gong, K., Lai, H., Zhu, J., Yin, J.: Soft-gated warping-gan for pose-guided person image synthesis. In: NeurIPS (2018)
9. Fu, C., Hu, Y., Wu, X., Wang, G., Zhang, Q., He, R.: High fidelity face manipulation with extreme pose and expression. arXiv preprint arXiv:1903.12003 (2019)
10. Ganin, Y., Kononenko, D., Sungatullina, D., Lempitsky, V.: Deepwarp: Photorealistic image resynthesis for gaze manipulation. In: ECCV (2016)
11. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR (2016)
12. Geng, J., Shao, T., Zheng, Y., Weng, Y., Zhou, K.: Warp-guided gans for single-photo facial animation. In: SIGGRAPH Asia 2018 Technical Papers (2018)
13. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing (2010)
14. Ha, S., Kersner, M., Kim, B., Seo, S., Kim, D.: Marionette: Few-shot face reenactment preserving identity of unseen targets. In: AAAI (2020)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a nash equilibrium. arXiv preprint arXiv:1706.08500 (2017)
16. Huang, R., Zhang, S., Li, T., He, R.: Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In: ICCV (2017)
17. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. arXiv preprint arXiv:1804.04732 (2018)
18. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: ICML (2017)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T., Van Knippenberg, A.: Presentation and validation of the radboud faces database. Cognition and emotion (2010)
21. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: ECCV (2018)

22. Li, M., Zuo, W., Zhang, D.: Deep identity-aware transfer of facial attributes. arXiv preprint arXiv:1610.05586 (2016)
23. Liao, J., Yao, Y., Yuan, L., Hua, G., Kang, S.B.: Visual attribute transfer through deep image analogy. ACM Trans. Graph. (2017)
24. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: NeurIPS (2017)
25. Liu, M.Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., Kautz, J.: Few-shot unsupervised image-to-image translation. In: ICCV (2019)
26. Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. In: NeurIPS (2016)
27. Liu, W., Piao, Z., Min, J., Luo, W., Ma, L., Gao, S.: Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In: ICCV (2019)
28. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: CVPR (2016)
29. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)
30. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. arXiv preprint arXiv:1712.00479 (2017)
31. Natsume, R., Yatagawa, T., Morishima, S.: Fsnet: An identity-aware generative model for image-based face swapping. In: ACCV (2018)
32. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NIPS-W (2017)
33. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: Anatomically-aware facial animation from a single image. In: ECCV (2018)
34. Qian, S., Lin, K.Y., Wu, W., Liu, Y., Wang, Q., Shen, F., Qian, C., He, R.: Make a face: Towards arbitrary high fidelity face manipulation. In: ICCV (2019)
35. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
36. Shu, Z., Sahasrabudhe, M., Alp Guler, R., Samaras, D., Paragios, N., Kokkinos, I.: Deforming autoencoders: Unsupervised disentangling of shape and appearance. In: ECCV (2018)
37. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. In: NeurIPS (2019)
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
39. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR (2016)
40. Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. arXiv preprint arXiv:1611.02200 (2016)
41. Tripathy, S., Kannala, J., Rahtu, E.: Icface: Interpretable and controllable face reenactment using gans. arXiv preprint arXiv:1904.01909 (2019)
42. Wang, T.C., Liu, M.Y., Tao, A., Liu, G., Kautz, J., Catanzaro, B.: Few-shot video-to-video synthesis. In: NeurIPS (2019)
43. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
44. Wiles, O., Sophia Koepke, A., Zisserman, A.: X2face: A network for controlling face generation using images, audio, and pose codes. In: ECCV (2018)
45. Wolf, L., Taigman, Y., Polyak, A.: Unsupervised creation of parameterized avatars. In: ICCV (2017)

46. Wu, R., Tao, X., Gu, X., Shen, X., Jia, J.: Attribute-driven spontaneous motion in unpaired image translation. In: ICCV (2019)
47. Wu, W., Cao, K., Li, C., Qian, C., Loy, C.C.: Transgaga: Geometry-aware unsupervised image-to-image translation. In: CVPR (2019)
48. Wu, W., Zhang, Y., Li, C., Qian, C., Change Loy, C.: Reenactgan: Learning to reenact faces via boundary transfer. In: ECCV (2018)
49. Yi, P., Wang, Z., Jiang, K., Jiang, J., Ma, J.: Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In: ICCV (2019)
50. Yi, Z., Zhang, H.R., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: ICCV (2017)
51. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. arXiv preprint arXiv:1905.08233 (2019)
52. Zhan, F., Zhu, H., Lu, S.: Spatial fusion gan for image synthesis. In: CVPR (2019)
53. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: ICML (2019)
54. Zhang, Y., Zhang, S., He, Y., Li, C., Loy, C.C., Liu, Z.: One-shot face reenactment. In: BMVC (2019)
55. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: AAAI (2019)
56. Zhou, T., Tulsiani, S., Sun, W., Malik, J., Efros, A.A.: View synthesis by appearance flow. In: ECCV (2016)
57. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)
58. Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: Progressive pose attention transfer for person image generation. In: CVPR (2019)