Pose Guided Image Generation from Misaligned Sources via Residual Flow Based Correction

Jiawei Lu[1], He Wang[2], Tianjia Shao[1*], Yin Yang[3], Kun Zhou[1]
[1] State Key Lab of CAD&CG, Zhejiang University
[2] University of Leeds
[3] Clemson University
lujiawei23@gmail.com, h.e.wang@leeds.ac.uk, tjshao@zju.edu.cn, yin5@clemson.edu, kunzhou@zju.edu.cn

Abstract

Generating new images with desired properties (e.g. new view/poses) from source images has been enthusiastically pursued recently, due to its wide range of potential applications. One way to ensure high-quality generation is to use multiple sources with complementary information such as different views of the same object. However, as source images are often misaligned due to the large disparities among the camera settings, strong assumptions have been made in the past with respect to the camera(s) or/and the object in interest, limiting the application of such techniques. Therefore, we propose a new general approach which models multiple types of variations among sources, such as view angles, poses, facial expressions, in a unified framework, so that it can be employed on datasets of vastly different nature. We verify our approach on a variety of data including humans bodies, faces, city scenes and 3D objects. Both the qualitative and quantitative results demonstrate the better performance of our method than the state of the art.

Introduction

Controlled image generation from source images is capable of generating scenes in unseen views and objects with pre-defined poses. It has a wide range of applications, e.g people with new poses (Ma et al. 2017), faces with new expressions (Zakharov et al. 2019) and scenes from different angles (Sun et al. 2018), and hence has attracted attention. The key challenge in such research is to recover the hidden information from sparse view points. One popular setting is to employ a single source image to generate new images with new poses/views. Despite recent successes (Ma et al. 2017; Siarohin et al. 2018; Zhang et al. 2021), ambiguity caused by the limited information available in a single image still makes it difficult to synthesize a high-quality image with large pose differences (e.g. generating the back view of a person when given only the front view). Consequently, high-quality generation is still an open challenge.

In theory, employing multiple source images with complementary information should mitigate the problem. However, in practice, this setting unfortunately brings additional challenges: the source images especially in-the-wild ones are not taken by calibrated cameras, leading to severe misalignment. Given the huge size of the camera space (possible camera poses), it is not straightforward to design a general solution. Consequently, strong assumptions have to be made. If the object is assumed to be rigid, multi-view images can help synthesize novel views (Sun et al. 2018; Zhou et al. 2016). If deformable objects are involved, certain types of alignment or transformations need to be assumed, such as feature averaging can help synthesize new facial expressions (Zakharov et al. 2019), but at the cost of losing the details of the source images; affine transformation can help synthesize new human poses (Lathuilière et al. 2020), but incapable of handling large pose deformation especially for non-rigid deformation such as clothes.

In this paper, we seek a general framework for controlled multi-source image generation. The framework takes as input multiple (misaligned) source images and source poses as well as a target pose, and predicts a new image under the target pose while keeping the source appearance. One key challenge is to impose parsimonious assumptions on the source images, so that they can differ in the camera pose, the camera-object distance, occlusions/lighting, etc. We aim to simultaneously deal with view/pose/expression variances and meanwhile synthesize high quality images with realistic details. One intuitive solution is to warp the features of each source image then fuse them for the target pose. However, two issues appear in such an approach. First, each source image is only a partial observation of the object, and all source images are misaligned. As a result, fusing such features inevitably leads to blurring in the target image. Second, the partial observability essentially dictates that different regions in a source image provides information with different levels of confidence (i.e. occluded areas having low confidence). Further, source images have different importance, so do their high/low confidence areas by association. This hierarchical structure of importance among source images cannot be captured via simple treatments, e.g. an occlusion map on a source image (Ren et al. 2020), or attention maps merely distinguishing the relative importance of different views (Sun et al. 2018).

To tackle the above challenges, we propose a novel fusion mechanism. Our framework adopts the state-of-the-art flow-based strategy, which first learns to warp each source feature to match the target pose at different levels, and then fuses

*Corresponding author.

Figure 1: Image generation on body poses (left) and facial expressions (right). Our model takes an arbitrary number of misaligned source images (I_1-I_4) and a target pose (P_t) to generate new images (Ours). I_t is the target ground truth.

these features in a decoder to synthesize the image. To tackle the challenge of hierarchical feature confidence, we propose to simultaneously predict the attention map and occlusion map for each source in the source feature extractor. The attention maps indicate the important source regions, and the occlusion maps dictate which part is invisible and should be inpainted. This way, the warped source features can be fused while being aware of the confidences of different source parts and invisible regions. To address the feature misalignment issue, we propose a novel residual-fusing (RF) block to correct the warping. A RF block consists of two modules: residual module and fusing module. The residual module corrects the warping of the source features and learns a residual flow for each warped source feature to match the fused feature from previous level. The fuse module takes the output of the residual module, and corrects the warped feature via an occlusion map. Then the corrected features are further fused by attention maps and sent to the next block. Overall, the RF blocks are repeated many times at multiple feature layers, so that different source features can be warped into a consistent space and be decoded to generate images with less artifacts and blurring.

Formally, our contributions include:

- a new general framework for controlled multi-source image generation, which can effectively capture view/pose/expression variances and synthesize high quality images with realistic details.
- a new Residual-Fusing block to systematically reconcile the conflicts caused by the misalignment of multiple (calibrated) sources.
- comprehensive experiments and comparisons on multiple distinctive datasets across different tasks to demonstrate the superiority of multi-source image generation under our general framework.

Related Work

Single Source Image Generation. Single source image generation aims to synthesize new images given a source image and a target pose. It involves many tasks including human pose transfer, novel view synthesis, facial image generation, etc. Ma et al. (2017) first introduced pose-guided human image generation and proposed a two-stage adversarial framework, which first synthesizes a coarse person image and then refines the result. Balakrishnan et al. (2018) presented a modular GAN network which decouples different body parts into layers and moves them by affine transformation. Siarohin et al. (2018) applied affine transformations in feature space by deformable skip connections. Zhu et al. (2019) proposed a novel block which simultaneously updates the pose code and appearance code in a coarse to fine manner. Although these works can synthesize correct global structures, they fail to preserve the local texture details provided in source images. In contrast, flow-based methods can better transfer the details such as clothing and texture. Han et al. (2019) proposed a three-stage network which first generates a semantic parsing map and then learns a flow of each semantic region. However, an extra refinement network is required as they predict the flow at the pixel level. Ren et al. (2020) presented a global flow and local attention architecture to generate vivid textures, but they struggled to synthesize unseen regions from a single source image.

Multi-source Image Generation. Our work is closely related to multi-source image generation methods. Zhou et al. (2016) utilized multiple views of an object to generate novel view given a target camera pose. They predict a pixel flow map together with a confidence map for each single view and merge them together by confidence maps. Sun et al. (2018) improved (Zhou et al. 2016) by adding a convolutional LSTM generator (Xingjian et al. 2015) to hallucinate the missing pixels from source view. Inspired by them, we also employ confidence maps and target generator in our work, but there are two main differences between our work

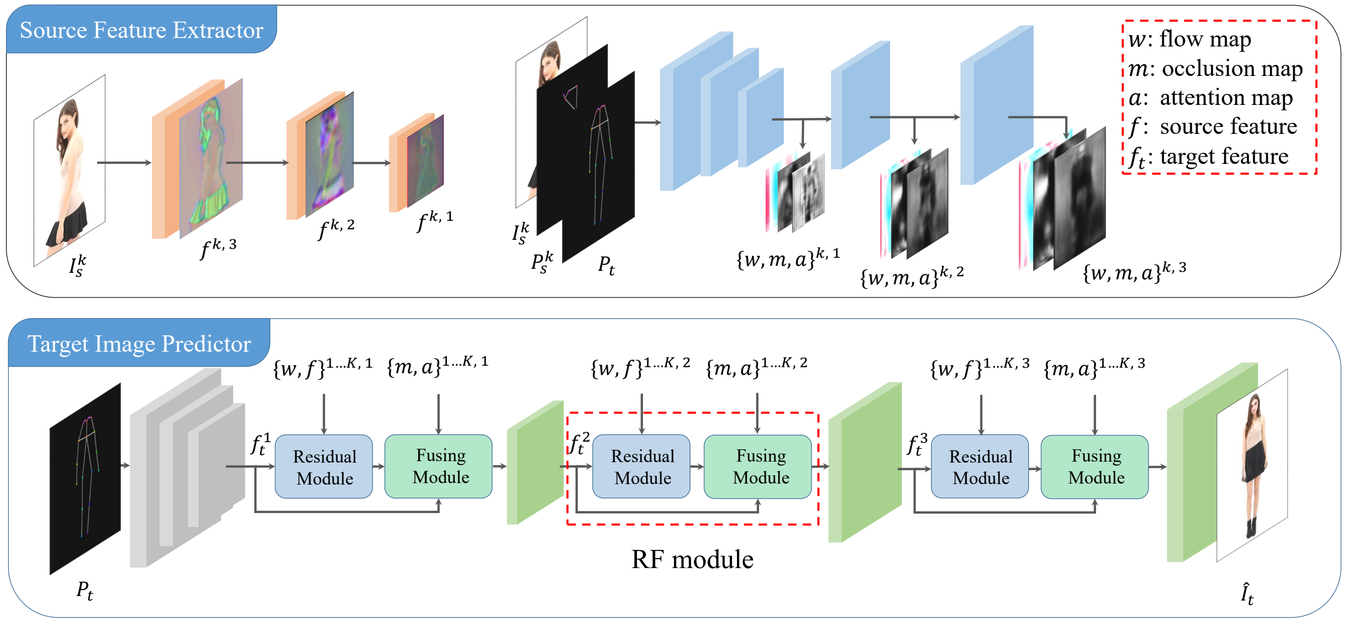

Figure 2: Network overview. Given K input images $\{I_s^k\}$, their poses $\{P_s^k\}$ and target pose P_t, we aim to generate a new image $\hat{I}_t$ in target pose P_t. Our source feature extractor first extracts the image features $\{f^k\}$ and generates flow maps $\{w^k\}$, occlusion maps $\{m^k\}$, attention maps $\{a^k\}$ at different levels for each source independently. The Target Image Predictor first encodes the target pose into target feature f_t^1. Then, the proposed Residual-Fusing Module in the decoder part of Target Image Predictor repeats at different feature levels to gradually fuse the sources and generate the target image $\hat{I}_t$.

and (Sun et al. 2018). First, Sun et al. (2018) conducted experiment mainly on rigid objects such as cars and chairs, while we can handle more general tasks, including highly non-rigid human image generation and facial image transfer. Second, they aggregated source images at pixel level directly, while we found it inappropriate in more general cases due to the complex texture and large motion, thus we propose to warp multi-level features instead of warping pixels. Lathuiliere et al (2020) introduced an attention-based decoder for multi-source human pose transfer. Some other works (Chan et al. 2019; Wang et al. 2018; Liu et al. 2019) had to train a model for each target, which limits their applications. Wang et al. (2019) improved (Wang et al. 2018) by generating network weights dynamically from reference images, however, continuous videos are required to calculate optical flow. In contrast, our approach is more general and can deal with flexible number of inputs with arbitrary poses.

Method

Overview

We propose a general generative adversarial network for multi-source pose guided image generation. Here the 'pose' can be any structural information of an image, e.g. human joints, view angles, facial landmarks, etc. Let K be the number of sources. Our generator G takes $\{I_s^k, P_s^k\}_{k=1,\ldots,K}$ and P_t as input, where I_s^k denotes the i-th input image, P_s^k denotes the corresponding pose representation, and P_t denotes the target pose. Our goal is to synthesize a new image $\hat{I}_t$ matching the target pose and meanwhile keep the source appearance. G can be written as:

$$\hat{I}_t = G(\{I_s^k\}, \{P_s^k\}, P_t). \tag{1}$$

Now we give a general overview of our architecture. Our model consists of two parts: source feature extractor and target image predictor. As shown in Fig. 2, in the source feature extractor, for each source, we estimate initial flow maps for warping the source features at different levels, and simultaneously predict the corresponding attention maps and occlusion maps. With the initial flow maps, the warped source features can be roughly aligned with the target pose at different levels. The necessity of multi-level modeling is primarily because misaligned features exist globally e.g. human poses, as well as locally e.g. textures on clothing. During the fusion, the attention maps play a role to select the important source regions among different sources, and the occlusion maps indicate which part is invisible and should be inpainted. As these source features are warped to match the target pose which only provides sparse structural information, directly fusing them will inevitably cause feature misalignment, leading to artifacts such as blurring and ghosting. Therefore, in the target image predictor, the residual-fusing (RF) module is brought up to further correct the warped features and fuse them. It contains two sub-modules: residual module and fusing module. At each feature level, the residual module takes the initially warped feature from each source branch, and learns a residual flow to further warp the feature to match the target feature from the previous level. The corrected source features are further sent to the fusing module, which performs a weighted aggregation of differ sources using the occlusion maps and attention maps, a outputs the fused target feature to the next level.

Source Feature Extractor

The source feature extractor F takes I_s^k, P_s^k, P_t as input a generates the initial flow field w^k, attention map a^k and clusion map m^k (as shown on the top right of Fig. 2):

$$w^k, a^k, m^k = F(I_s^k, P_s^k, P_t),$$

where w^k stores the coordinate displacements between source and the target features, and a^k and m^k has co tinuous values between 0 and 1. m^k measures how target feature is visible in a source at a certain positi and a^k indicates which source is more relevant to the t get at a certain position. We design F as a fully convolutional network with a pyramid architecture, which outputs w^k, a^k and m^k at N different resolutions, i.e. $w^k, a^k, m^k = \{w^{k,i}, a^{k,i}, m^{k,i}\}, i = 1...N$. $w^{k,i}, a^{k,i}$ and $m^{k,i}$ share the same backbone of F except their output layers. Source image feature f^k is extracted by another convolutional network(shown on the top left of Fig. 2). Please refer to the supplementary material for details. The attention map and occlusion map will be jointly applied in the subsequent Fusing Module to ensure a globally consistent feature fusion.

Target Image Predictor

After feature extraction, the image prediction is handled in the target image predictor. The major difficulty here is to fuse the source features into one consistent target feature and meanwhile reduce the feature misalignment from the initial warping. As shown on the bottom of Fig. 2, the target image generator starts from the target pose P_t and then goes through several down-sampling layers to get the initial target feature f_t^1. Then, at each feature level i in the predictor, a Residual-Fusing (RF) block is deployed to correct the warping field of each source feature based on the target feature from the previous level, and then fuse different sources together to output the target feature to the next level $i+1$.

The RF block repeats at different feature levels and has K branches to deal with K different sources. It can be further divided into two sub-modules, named residual module and fusing module (as shown in Fig. 3). Now we take feature level i as example and give a detailed description of each sub-module.

Residual Module. The source feature extractor provides a coarse flow field to warp the source to match the target pose, which only provides sparse structural information. Directly fusing the coarsely warped features inevitable causes misalignment and artifacts. The residual module gives the network the ability to further correct the initial flow, by learning a residual flow from the initially warped source feature and the fused feature from previous level.

At feature level i, the k-th residual module receives the source image feature $f^{k,i}$, and the flow map $w^{k,i}$ from the k-th source feature extractor, together with the fused feature f_t^i from previous level. We first warp the source feature $f^{k,i}$

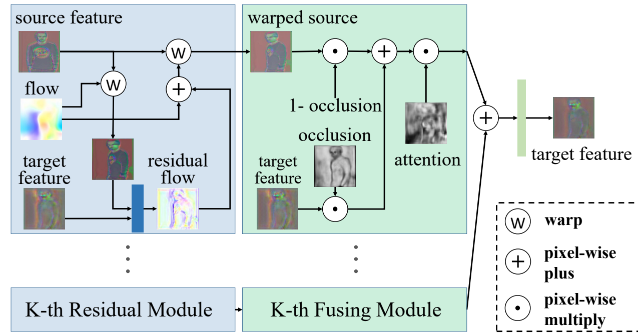

Figure 3: The residual-fusing(RF) module repeats at different feature levels and has K same branches. For a specific branch at level i, the residual module computes the residual flow to further warp the source feature. Then, the fusing module performs feature matting between the warped source feature and previous target feature by the occlusion map. Finally, K features are fused together by the attention maps and output to the next level. Zoom in for better details.

with the initial flow map $w^{k,i}$ to the target pose by:

$$f_w^{k,i} = \mathcal{W}(f^{k,i}, w^{k,i}). \tag{3}$$

Then we predict a residual flow $r^{k,i}$ by a residual flow predictor R_i from f_t^i and $f_w^{k,i}$:

$$r^{k,i} = R^i(f_t^i, f_w^{k,i}). \tag{4}$$

Note that R^i is designed to share weights across different branches, making it possible for the model to accept arbitrary number of inputs at inference time.

With the learned residual flow, we get the refined flow field by adding the residual flow to the initial flow. We then warp each source feature $f^{k,i}$ by the refined flow and get the refined warped feature $\hat{f}_w^{k,i}$ of source branch k:

$$\hat{f}_w^{k,i} = \mathcal{W}(f^{k,i}, w^{k,i} + r^{k,i}). \tag{5}$$

Fusing Module. With the refined warped source feature $\hat{f}_w^{k,i}$, the fusing module merges $\hat{f}_w^{k,i}$ and the previous fused target feature f_t^i using the occlusion map $m^{k,i}$. Then, these merged features of K branches are fused together into one, by a weighted summation over the K attention maps $\{a^{k,i}\}$, where $a^{k,i}$ are normalized by a softmax operation at pixel level to stabilize the gradient during training. Finally, the fused feature goes through a decode layer D^i to output the next level target feature f_t^{i+1}:

$$f_t^{i+1} = D^i(\sum_{i=1}^{K} a^{k,i} \cdot (\hat{f}_w^{k,i} \cdot (1 - m^{k,i}) + f_t^i \cdot m^{k,i})) \tag{6}$$

Training

We train our model in two stages. First, without the ground truth flow field, we warm up the flow generator F using the sample correctness loss (Ren et al. 2019). We also take the regularization loss (Ren et al. 2020) to constrain the smoothness of the flow.

$$\mathcal{L}_{flow} = \lambda_{cor}\mathcal{L}_{cor} + \lambda_{reg}\mathcal{L}_{reg} \tag{7}$$

where the sampling correctness loss $\mathcal{L}_{cor}$ maximizes the cosine similarity between the VGG features of the warped source and target and force the flow field $w^{k,i}$ to sample the similar regions. The regularization term $\mathcal{L}_{reg}$ penalize the local regions where the transformation is not an affine transformation. Then, with the pre-trained flow generator, we train our full model in an end-to-end manner. The full loss can be defined as:

$$\mathcal{L} = \mathcal{L}_{flow} + \mathcal{L}_{con} + \mathcal{L}_{adv} \tag{8}$$

where $\mathcal{L}_{con}$ is a content loss, and $\mathcal{L}_{con} = \lambda_{l_1}\mathcal{L}_{l_1} + \lambda_{per}\mathcal{L}_{per} + \lambda_{sty}\mathcal{L}_{sty}$. $\mathcal{L}_{l_1}$ minimizes the L_1 distance of the generated image and target image, $\mathcal{L}_{l_1} = ||\hat{I}_t - I_t||_1$. $\mathcal{L}_{per}$ and $\mathcal{L}_{sty}$ are inspired by (Johnson, Alahi, and Fei-Fei 2016). The perceptual loss $\mathcal{L}_{per}$ aims to penalize the L_1 distance between features extracted from specific layers of a pre-traind VGG network:

$$\mathcal{L}_{per} = \sum_i ||\phi_i(\hat{I}_t) - \phi_i(I_t)||_1 \tag{9}$$

where ϕ_i denotes the i-th layer of the VGG-19 network. The style loss $\mathcal{L}_{sty}$ uses the Gram matrix of VGG features to maximize the style similarity between the images:

$$\mathcal{L}_{sty} = \sum_j ||G_j^{\phi}(\hat{I}_t) - G_j^{\phi}(I_t)||_1 \tag{10}$$

where G_j^{ϕ} denotes the Gram matrix calculated from ϕ_j. Finally, we use a standard adversarial loss $\mathcal{L}_{adv}$:

$$\mathcal{L}_{adv} = \mathbb{E}[log(1-D(G(\{I_s^k\}, \{P_s^k\}, P_t)))] + \mathbb{E}[log(D(I_t))] \tag{11}$$

Implementation Details

We implement our model using PyTorch (Paszke et al. 2019) framework on a PC with four NVIDIA GTX 2080Ti GPUs. We adopt the Adam optimizer($\beta_1 = 0.9, \beta_2 = 0.999$) with a learning rate of 0.0001. The batch size is fixed to 5 for all tasks except Market-1501, in which batch size is set to 8. For more details, please refer to the supplementary material.

Experiments

Datasets. Since our model is designed to be general, we conduct experiments on three different tasks including pose transfer, view synthesis, and facial expression transfer on five challenging datasets. For human pose transfer, we use DeepFashion In-shop Clothes Retrieval Benchmark (Liu et al. 2016) and person re-identification dataset Market-1501 (Zheng et al. 2015). For novel view synthesis, we use real-world scenes (KITTI Visual Odometry Dataset (Geiger, Lenz, and Urtasun 2012)) and rendered objects (ShapeNet chair dataset (Chang et al. 2015)). For facial expression transfer, we use the talking videos dataset Voxceleb2(Chung, Nagrani, and Zisserman 2018). More details on the dataset can be found in the supplementary material.

These tasks are challenging in different ways. Human pose transfer needs to handle deformable human bodies with full and partial views, along with details on clothing; view synthesis needs to consider complex image semantics and features such as shadows; facial expression transfer needs to model consistent and realistic facial features. To handle them under one method, they show the generality of our model.

Metrics. How to evaluate the generated images remains an open problem in generative models. We follow (Ren et al. 2020; Zhang et al. 2021) and calculate the Frechet Inception Distance (FID) (Heusel et al. 2017) and Learned Perceptual Image Patch Similarity(LPIPS)(Zhang et al. 2018) to evaluate the performance of our model. For Market-1501, we further report the Mask-LPIPS (MLPIPS) score proposed in (Ma et al. 2017) to exclude the influence of the background. Besides, we perform a user study to evaluate the visual quality of the generated images.

Qualitative Results

We first show our results on the DeepFashion Benchmark with two input images in Fig. 1 Left. In all cases, the target contains a novel view and pose, which means the model has to learn to correctly align feature of different sources. Further, the clothing details are also transferred well (e.g. the shirts in row 1, the hat in row 2,3), thanks to our multi-level feature modeling. Fig. 1 Right shows the results on facial expression transfer on Voxceleb2. Realistic unseen expressions are generated by our model and source identities are preserved (row 1). Face in new head poses can also be generated. More results on KITTI and ShapeNet can be found in the supplementary material.

Comparisons

We compare our method with a variety of baseline methods across all tasks. For the human pose transfer task, we compare our approach with Def-GAN (Siarohin et al. 2018), PATN (Zhu et al. 2019), GFLA (Ren et al. 2020), PISE (Zhang et al. 2021), ADG (Men et al. 2020), ABF (Lathuilière et al. 2020). All baselines except ABF are single source based methods. For ABF, we train and evaluate their model using the same train/test split. For the other baseline methods, we use the pre-trained models and evaluate the performance on the testing set directly. For the novel view synthesis task, we compare our method with M2N (Sun et al. 2018). We run the pre-trained model offered by the author to get results on KITTI and ShapeNet chair dataset. For the face generation task, we compare our method with NH-FF(the feed forward result of (Zakharov et al. 2019)). We implement their method and report results using the same train/test split of Voxceleb2.

For the human image generation task, our method outperforms all baseline methods on all the metrics on both single-source and multi-source settings by a large margin, shown in Table 1. The LPIPS scores drop significantly when more source images are used, which demonstrates the effectiveness of utilizing multiple sources. Our model also shows superiority on other datasets, as shown in Table 2, 3.

The qualitative results on the DeepFashion dataset are shown in Fig. 4. The baseline methods (Zhu et al. 2019; Siarohin et al. 2018; Men et al. 2020; Zhang et al. 2021; Lathuilière et al. 2020) fail to keep the source appearance

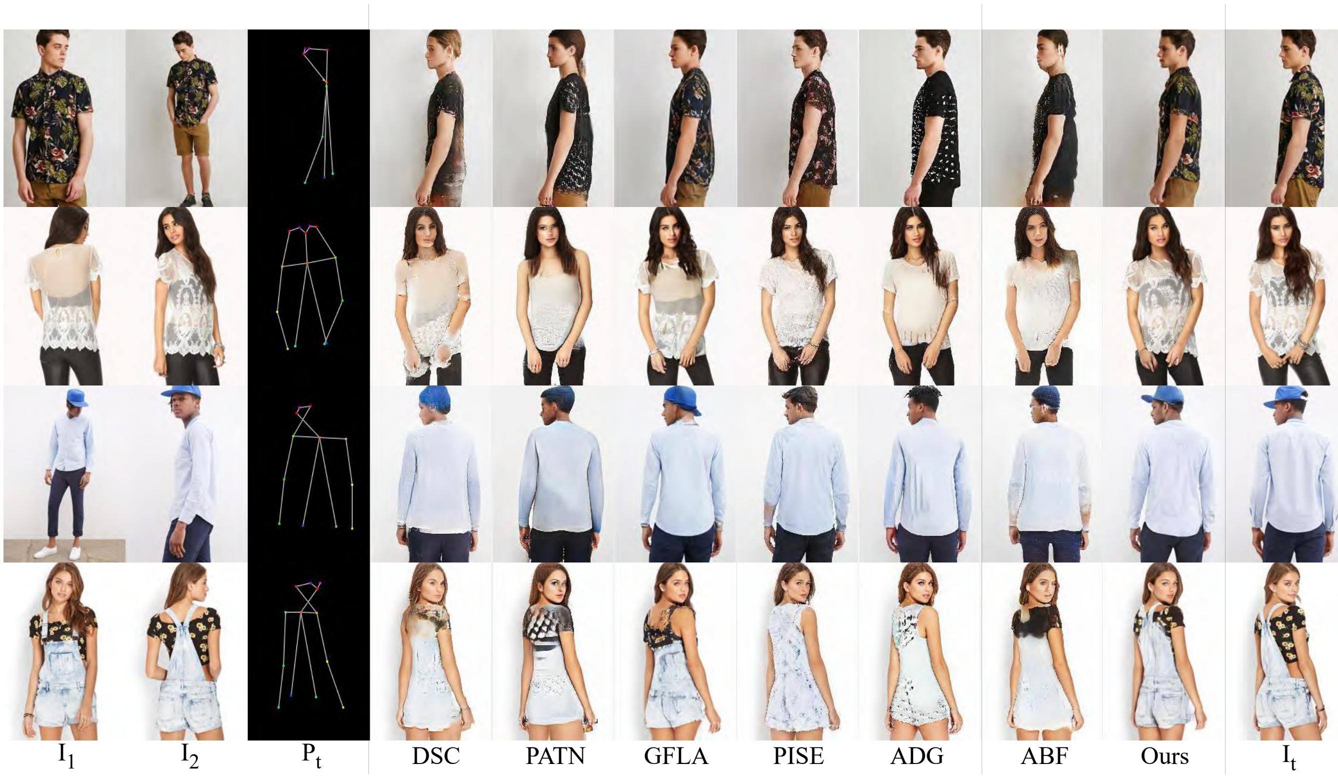

Figure 4: Qualitative comparison on the DeepFashion dataset. DSC (Siarohin et al. 2018), PATN (Zhu et al. 2019), GFLA (Ren et al. 2020), PISE (Zhang et al. 2021), ADG (Men et al. 2020) are single source based methods, in which only I_1 is used as input. ABF (Lathuilière et al. 2020) and Ours are multi-source based methods, in which I_1 and I_2 are used as input.

		DeepFashion		Market-1501		
K	Model	FID	LPIPS	FID	LPIPS	MLPIPS
1	DSC	21.542	0.2384	24.861	0.2984	0.1495
	PATN	20.632	0.2553	22.753	0.3181	0.1585
	GFLA	<u>9.872</u>	<u>0.1963</u>	<u>19.750</u>	<u>0.2817</u>	0.1483
	ADG	14.476	0.2253	-	-	-
	PISE	11.524	0.2077	-	-	-
	ABF	27.303	0.2753	32.588	0.3015	<u>0.1480</u>
	Ours	**9.750**	**0.1867**	**17.362**	**0.2730**	**0.1418**
2	ABF	23.529	0.2577	30.274	0.2878	0.1390
	Ours	**10.135**	**0.1766**	**15.716**	**0.2668**	**0.1339**
3	ABF	26.759	0.2376	33.270	0.2870	0.1314
	Ours	**12.785**	**0.1689**	**16.263**	**0.2604**	**0.1277**

Table 1: Quantitative comparison with SOTA methods on the DeepFashion dataset and the Market-1501 dataset.

when the clothes pattern is complex (e.g. they fail to capture the texture details of the clothes in row 1). Flow-based method (Ren et al. 2020) can preserve the details of the source, but struggles when there is a big gap between the source pose and the target one(e.g. It fails to generate correct back view from the front view in row 2, and predicts a wrong hat direction in row 3). (Lathuilière et al. 2020) utilizes multiple inputs to generate the target view but fails to maintain source details in the generated images. In contrast, our

		Chair		KITTI	
K	Model	FID	LPIPS	FID	LPIPS
2	M2N	28.876	0.1155	18.798	0.1958
	Ours	**10.123**	**0.09607**	**8.505**	**0.1721**
4	M2N	21.920	0.0901	-	-
	Ours	**7.697**	**0.0729**	-	-

Table 2: Quantitative comparison with M2N (Sun et al. 2018) on the KITTI dataset and the ShapeNet chair dataset. Our method gets lower FID scores and LPIPS scores than M2N on both datasets.

model transfers high-fidelity details from the source images, and extracts information from different sources to overcome the single source ambiguity.

Qualitative comparisons on other datasets (Voxceleb2, KITTI and ShapeNet chair) can be found in the supplementary material.

Ablation Study

We present an ablation study on the human pose transfer task to clarify the impact of each part of our proposed method.

Baseline. Our baseline model is U-Net architecture with feature warping, with no residual flow, attention map or occlusion map. Source features are fused by averaging.

Without occlusion (w/o. occ). This model is designed to

K	Model	FID	LPIPS
2	NH-FF	37.266	0.3150
	Ours	**7.100**	**0.2130**
4	NH-FF	37.457	0.3131
	Ours	**7.690**	**0.2084**

Table 3: Quantitative comparison with NH-FF (Zakharov et al. 2019) on the Voxceleb2 dataset. Our method outperforms NH-FF in both metrics.

		DeepFashion		Market-1501		
K	Model	FID	LPIPS	FID	LPIPS	MLPIPS
2	Baseline	11.078	0.1857	15.298	0.2714	0.1393
	w/o attn	10.835	0.1774	15.820	0.2676	0.1368
	w/o occ	10.421	0.1830	**15.249**	0.2689	0.1349
	w/o res	10.292	0.1777	17.701	0.2679	0.1483
	full	**10.135**	**0.1766**	15.716	**0.2668**	**0.1339**
3	w/o res	13.306	0.1710	18.092	0.2645	0.1302
	full	**12.785**	**0.1679**	**16.263**	**0.2604**	**0.1277**

Table 4: Ablation Study on the DeepFashion dataset and Market-1501 dataset.

see if the occlusion maps can benefits the learning. We remove the occlusion maps and source features are aggregated by attention maps.

Without attention (w/o. attn). The model is designed to see if the attention maps are effective in modeling confidence. We replace the attention maps of different views with the same value. The occlusion mechanism is adopted.

Without residual flow (w/o. res). In this configuration, we remove the residual block in the decoder to evaluate its contribution. Attention and occlusion mechanisms are employed in this model.

Qualitative results are shown in Fig. 5. The baseline method struggles when the source images have large differences in poses and views (e.g. the inner clothes in row 1, the dress in row 2), as it simply performs a weighted average over source features, without considering the relevant importance of each source. The baseline model also suffers from generating occluded parts and wrongly warped areas (e.g. face/arms of the man in row 5, hands of the woman in row 6). By adding the occlusion mechanism, the model can get improvements in these areas, but the model without attention mechanism tends to generate ghosting effects when the two sources are similar but at different scales (e.g. the collar of the man in row 4). The model with attention maps but without occlusion maps could also fail when the model synthesizes new contents (e.g. the the hand of the model in row 3, the face of the model in row 5,). And for the model without residual flow, the synthesized results suffer from blurry textures(the white spot in row 3, the vest in row 7) and irregular boundary of cloths(the boundary of the sweater in row 6). Detailed quantitative results are shown in Table 4.

User Study

We also conduct a user study to assess the visual quality. For each dataset, 30 volunteers are asked to accomplish two tasks: the first is a 'real or generated' test, following the protocol in (Ma et al. 2017; Siarohin et al. 2018). For each model, volunteers are shown 55 real and 55 generated images in a random order. The volunteers are asked to judge whether the displayed image is real or generated in one second. The other is a comparison task. The volunteers are asked to finish 55 questions on each dataset, each question containing image pairs generated by ours method and a baseline method, with the same source images and target pose. The volunteers are asked to choose the one with better quality. All samples are randomly selected. Overall, our method significantly outperforms the baseline methods. Detailed results are shown in the supplementary material.

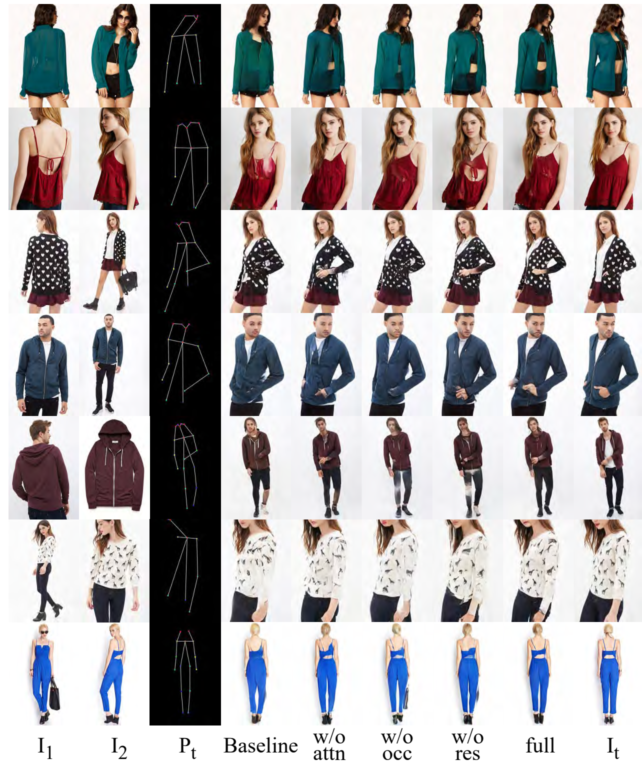

Figure 5: Ablation study. The first two columns show the inputs, the third column shows the target pose and the last column shows the target image. Results of different ablation models are shown in the middle. Zoom in for better details.

Conclusion

We have proposed a new general method for multi-source image generation. Given a guiding pose, our framework effectively rectifies the issues caused by the misalignment among the sources, which makes it widely applicable to datasets with in-the-wild images taken by un-calibrated cameras. The model generality has been tested on a variety of vastly different datasets including human poses, street scenes, faces and 3D objects, and verified by its universal successes. In exhaustive comparisons, our model outperforms the state-of-the-art methods in various tasks.

Acknowledgements

We thank the reviewers for their comments and suggestions for improving the paper. The work was supported by NSF China (No. 61772462, No. U1736217, No. 61772457), the 100 Talents Program of Zhejiang University, and NSF (2016414 and 2011471).

References

Balakrishnan, G.; Zhao, A.; Dalca, A. V.; Durand, F.; and Guttag, J. 2018. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8340–8348.

Chan, C.; Ginosar, S.; Zhou, T.; and Efros, A. A. 2019. Everybody Dance Now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Chung, J. S.; Nagrani, A.; and Zisserman, A. 2018. VoxCeleb2: Deep Speaker Recognition. In *INTERSPEECH*.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Han, X.; Hu, X.; Huang, W.; and Scott, M. R. 2019. ClothFlow: A Flow-Based Model for Clothed Person Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*.

Lathuilière, S.; Sangineto, E.; Siarohin, A.; and Sebe, N. 2020. Attention-based fusion for multi-source human image generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 439–448.

Liu, L.; Xu, W.; Zollhoefer, M.; Kim, H.; Bernard, F.; Habermann, M.; Wang, W.; and Theobalt, C. 2019. Neural rendering and reenactment of human actor videos. *ACM Transactions on Graphics (TOG)*, 38(5): 1–14.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ma, L.; Jia, X.; Sun, Q.; Schiele, B.; Tuytelaars, T.; and Van Gool, L. 2017. Pose Guided Person Image Generation. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Men, Y.; Mao, Y.; Jiang, Y.; Ma, W.-Y.; and Lian, Z. 2020. Controllable Person Image Synthesis with Attribute-Decomposed GAN. In *Computer Vision and Pattern Recognition (CVPR), 2020 IEEE Conference on*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32: 8026–8037.

Ren, Y.; Yu, X.; Chen, J.; Li, T. H.; and Li, G. 2020. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7690–7699.

Ren, Y.; Yu, X.; Zhang, R.; Li, T. H.; Liu, S.; and Li, G. 2019. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 181–190.

Siarohin, A.; Sangineto, E.; Lathuilière, S.; and Sebe, N. 2018. Deformable GANs for Pose-Based Human Image Generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sun, S.-H.; Huh, M.; Liao, Y.-H.; Zhang, N.; and Lim, J. J. 2018. Multi-view to Novel View: Synthesizing Novel Views with Self-Learned Confidence. In *European Conference on Computer Vision*.

Wang, T.-C.; Liu, M.-Y.; Tao, A.; Liu, G.; Kautz, J.; and Catanzaro, B. 2019. Few-shot Video-to-Video Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Liu, G.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. Video-to-Video Synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Xingjian, S.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, 802–810.

Zakharov, E.; Shysheya, A.; Burkov, E.; and Lempitsky, V. 2019. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhang, J.; Li, K.; Lai, Y.-K.; and Yang, J. 2021. PISE: Person Image Synthesis and Editing with Decoupled GAN. *arXiv preprint arXiv:2103.04023*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable Person Re-identification: A Benchmark. In *Computer Vision, IEEE International Conference on*.

Zhou, T.; Tulsiani, S.; Sun, W.; Malik, J.; and Efros, A. A. 2016. View Synthesis by Appearance Flow. In *European Conference on Computer Vision*.

Zhu, Z.; Huang, T.; Shi, B.; Yu, M.; Wang, B.; and Bai, X. 2019. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2347–2356.