General Object Pose Transformation Network from Unpaired Data

Yukun Su[1,2], Guosheng Lin[2*], Ruizhou Sun[1], and Qingyao Wu[1,3*]

[1] School of Software Engineering, South China University of Technology
[2] Nanyang Technological University, Singapore
[3] Pazhou Lab, Guangzhou, China
suyukun666@gmail.com

Abstract. Object pose transformation is a challenging task. Yet, most existing pose transformation networks only focus on synthesizing humans. These methods either rely on the keypoints information or rely on the manual annotations of the paired target pose images for training. However, collecting such paired data is laboring and the cue of keypoints is inapplicable to general objects. In this paper, we address a problem of novel general object pose transformation from unpaired data. Given a source image of an object that provides appearance information and a desired pose image as reference in the absence of paired examples, we produce a depiction of the object in that specified pose, retaining the appearance of both the object and background. Specifically, to preserve the source information, we propose an adversarial network with **S**patial-**S**tructural (SS) block and **T**exture-**S**tyle-**C**olor (TSC) block after the correlation matching module that facilitates the output to be semantically corresponding to the target pose image while contextually related to the source image. In addition, we can extend our network to complete multi-object and cross-category pose transformation. Extensive experiments demonstrate the effectiveness of our method which can create more realistic images when compared to those of recent approaches in terms of image quality. Moreover, we show the practicality of our method for several applications.

Keywords: Pose Transformation, Adversarial network, Semantically, Contextually

1 Introduction

Image-to-image translation tasks include image colorization [4], image super-resolution [20,64], style transfer [13], domain adaptation [35] and pose transformation [31,51], *etc.* Among them, we are interested in pose transformation, which has huge potential applications in re-enactment, character animation, movie or game making and so on. However, most recent approaches [2,32,23] merely explore human pose transformation, and such methods require abundant keypoints

* Corresponding authors.

Fig. 1. Illustrative examples of different general objects pose transformation. Given the desired pose image (1^{st} row) and the appearance image (2^{nd} row) in the absence of paired examples, we produce the output image (3^{rd} row) in that pose and retain the appearance of object and background. We can obtain high-quality images and apply the network to different object posture modalities. The generated samples are not cherry-picked, more samples are provided in supplementary material.

information [5,55] or paired data, *e.g.*, they collect the same person of different target poses for training. With these in mind, we argue that the previous works suffer from some **limitations**: (1) In addition to human, some other general objects should also be able to conduct pose transformation, which is helpful for wider applications. (2) As for the general objects, such human keypoints [5] and body mesh [15] information will not be suitable. (3) In real life, it is difficult for us to collect different postures of the same object, which is laborious and time-costly.

To address the issues mentioned above, we propose a **U**nified **F**ramework for general **O**bject **P**ose **T**ransformation with unpaired data, termed as **UFO-PT**. As shown in Fig 1, given the unpaired images that provide pose and appearance information, respectively, we can yield the output images in that pose while keeping the appearance of objects and background unchanged. Our method can be applied not only to human body pose transformation, but also to non-rigid objects such as mammals (*i.e.*, cow, sheep, horse, *etc.*) and birds, and even rigid objects such as vehicles.

In this paper, we propose a network which comprises four sub-blocks as shown in Fig 2: (1) The correlation matching block is introduced to align the unpaired images and warp the appearance image into the target pose. Specifically, we estimate two types of warpings inspired by [66] in different level: (i) Dense warping. (ii) Thin Plate Spline (TPS) warping [59]. The former has a high degree of freedom, which can be utilized to map pixels to be well-aligned with the target pose. While the latter roughly transfers the images into the desired pose but with well-preserved details, which can be utilized to retain the appearance information. (2) The **S**patial-**S**tructural (SS) block employs the information from the output of dense warping in the form of spatially-variant de-normalization [43] to progressively inject the spatial details to the generated network. (3) The

Texture-**S**tyle-**C**olor (TSC) block employs the information from the output of TPS warping to preserve the appearance details when synthesizing the results. (4) The generating block is responsible to combine the transformed foreground object and background to produce the output, which is semantically aligned to the target pose image while contextually related to the source appearance image. Moreover, our proposed method can be applied to some practical applications such as data augmentation and video imitation. Our contribution can be summarized as follows:

- We address the problem of general object pose transformation and propose a unified framework with unpaired data, which to our best knowledge, has not been well explored.
- With the proposed four sub-blocks in the network, we can generate more realistic transformed images in the desired pose preserving the original appearance and background compared to recent methods.
- Quantitative comparisons against several prior methods demonstrate the superiority of our approach, which can also be applied to several practical applications.

2 Related Work

Pose-Guided Human Image Generation: Skeletal pose cues [5,54,53] provide strong information and most previous methods are based on conditioned generative adversarial networks (CGAN) [34]. Ma *et al.* [31] generates human images conditioned on pose utilizing a two-stage network. Siarohin *et al.* [51] introduces deformable skip connections to spatially transform the features. Si *et al.* [49] proposes a multistage adversarial loss and separately generates the foreground and background. Zhu *et al.* [71] designs a progressive pose-attention transfer block to avoid the issues of capturing the complex structure of the global manifold. Some other works like [2,32,49,38,50] all combine target image along with source pose (2D keypoints) as inputs or use video optical flow [7] information to generate images by GANs. Liu *et al.* [23] later suggests to use SMPL [27] to disentangle the pose and shape, which can help to promote the transformed results. However, the above mentioned methods only focus on synthesizing humans, and they require paired data and keypoints information for training. It is difficult to collect such data and they fail to conduct general objects pose transformation, which will weaken their practicality. These shortcomings might limit their wide applications.

View Synthesis: View synthesis is a task in computer vision in which unseen camera views or poses of objects are synthesized given a prior image. Most view synthesis work has focused on simple rigid objects such as cars [14,19,42,44,69]. These methods rely on camera viewpoints and underlying 3D models. Recently, HoloGan [41], Graf [47] and π-GAN [6] propose to correctly inject 3D priors into the GAN framework to transform the 3D pose of 2D objects, while our proposed technique in this paper treats the problem as a 2D one and attempts to replace

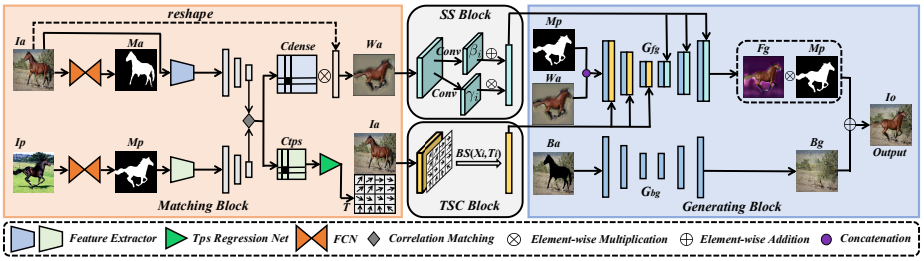

Fig. 2. The Overview of the UFO-PT architecture. For simplicity, we take "*horse*" as an example for input. Given the unpaired images I_a and I_p providing appearance and pose information, the matching block first aligns them and establishes the two types of warpings. Then, the generating block yield the transformed output I_o based on the different injected warped intermediates from SS block and TSC block, respectively.

pixels of one object with another one. Lv *et al.* [29] later addresses the problem of novel view synthesis for vehicles without exploiting additional 3D details but using stack hourglass [39] to obtain keypoint polygon. Likewise, some of these methods depend on the paired target pose data for training and the objects they work with are relatively easy and have a simple background.

Example-Guided Image Synthesis: Recently, a few works [30,3] propose to synthesize photorealistic images from semantic layout under the guidance of exemplars. Wang *et al.* [60] and Zhang *et al.* [68] both can readily be applied to human pose transformation that semantically consistent to the label maps. However, these methods require to constitute style consistency image pairs or generate images from abstract semantic label maps such as pixel-wise segmentation maps or sparse landmarks, which makes it unsuitable for general image translation and it is difficult to obtain instance-level labels.

Content-Style Image Translation: Unpaired image-to-image translation aims to map an image from a source domain to a target domain. Such methods as [17,70,16] encourage the translated domain to be faithfully reconstructed when mapping back to the original domain with cycle loss. Lorenz *et al.* [28] proposes an part-based disentangling method for object shape and TransGaGa [65] introduces geometry-aware technique for image translation. However, they either focus on human animation or object faces. More recently, Liu *et al.* [21] and Saito *et al.* [46] introduce more powerful methods to preserve the structure of the input image while emulating the appearance of the unseen domain. However, these methods fail to delicately control the output since this content-style translation will break the global information of the image in the pose transformation task.

3 Method

Correspondence Matching Block. To synthesize the transfer results, one of the main challenges is to establish the correlation between the input I_a and I_p. Inspired by [59] in the Virtual Try-on field, a good practice to facilitate the generation is to utilize warping methods to align the appearance image with the target pose image first before feeding them into the generating network. However, we just want to warp the foreground objects to preserve the local style but retain the global background details.

To this end, as shown in Fig 2, we first adopt an off-the-shelf unsupervised salient object detection network [40] to obtain the mask M_a and M_p. Then, we employ two separate feature extractors F_A and F_B to extract high-level features f_a and f_p of I_a and M_p, where $f_a = F_A(I_a, \theta_a)$ and $f_p = F_B(M_p, \theta_p)$. The merit of this is that when conducting warping, it only pays attention to the foreground objects and will not be interfered by the background. After that, we estimate the correspondence matrices $C_{dense} \in \mathbb{R}^{\frac{HW}{4} \times \frac{HW}{4}}$ using a sliding kennel size = 1, stride = 1 and padding = 0, while $C_{tps} \in \mathbb{R}^{\frac{HW}{16} \times \frac{HW}{16}}$ by utilizing a sliding kennel size = 4, stride = 4 and padding = 0 for spatial reduction, where H and W indicate the spatial size of the original input image.

As for dense correspondence warping, we propose to match the features of f_a and f_p by using cosine similarity as follows:

$$C_{dense} = \frac{(f_a - \mu_a)^T (f_p - \mu_p)}{||f_a - \mu_a|| ||f_p - \mu_p||}, \tag{1}$$

where μ_a and μ_p represent the mean vectors. We then calculate the weighted average to estimate the dense correspondence warping in the form as [68]:

$$W_a = \sum \text{Softmax}(\alpha C_{dense} \cdot I_a, \text{dim} = 1), \tag{2}$$

where α is a hyper-parameter that controls the sharpness of the softmax function. To force the network to learn a reasonable dense semantic warping, we introduce a geometric loss as follows:

$$\mathcal{L}_{geo} = ||I_p - W_a||_1. \tag{3}$$

Although dense warping is capable to handle high degree of geometric changes, it fails to preserve the detailed style and texture information. To tackle this drawback, we further involve TPS warping, which can roughly transform the objects but with little information loss.

As for TPS warping, after obtaining C_{tps} matrix like C_{dense}, then we employ a regression net [59] to predict the corresponding control points and calculate the flow parameters T. Concretely, we use the following loss to restrict the transformation flow:

$$\mathcal{L}_{tps} = ||BS(I_a, T) - I_p||_1 + \mathcal{L}_{cst}, \tag{4}$$

where BS indicates the bilinear sampler operation. $\mathcal{L}_{cst}$ is a constraint loss [67] that restricts the warp distance and amplitude to prevent the internal patterns from losing natural information.

SS and TSC Block. (i) SS block utilizes the information from the output of the dense warping. Specifically, we employ the spatially-adaptive denormalization block [43] to project the spatially variant style to different activation decoder layers in the generating network as shown in Fig 2. Formally, let $F^i \in \mathbb{R}^{h \times w}$ denote the activations of the i-th layer of a deep convolutional network, we inject the dense warping information as follows:

$$\hat{F}^i = \gamma^i W_a \times \frac{F^i - \mu^i}{\sigma^i} + \beta^i W_a, \tag{5}$$

where $\sigma^i = \sqrt{\frac{1}{nhw}\sum((F^i)^2 - (\mu^i)^2)}$ and $\mu^i = \frac{1}{nhw}\sum F^i$, where n is batch sample number. We implement the functions γ^i and β^i by using a simple two-layer convolutional network. **(ii)** TSC block employs the information from the output of TPS warping and project the appearance details to different activation encoder layers in the generating network. Formally, let X^i denote the activations of the i-th layer of the network, we inject the TPS warping information as follows:

$$\hat{X}^i = \varphi^i(BS(I_a, T)) + X^i, \tag{6}$$

where we use a simple plain convolutional layer to obtain φ^i.

Generating Block. For G_{Fg}, we combine W_a and M_p as input. For G_{Bg}, we take $B_a = I_a \otimes (1 - M_a)$ as input, and they do not share parameters. The final output can be obtained as: $I_o = Fg \otimes M_p + Bg$. More details about network architectures are provided in supplementary material.

End-to-end Training. To encourage the training of different blocks benefit from each other, we train our model in a joint style, and we combine several different losses to produce high-quality transferred output images:

Perceptual-Loss: the final output should be semantically consistent with the desired pose image, we then minimize the semantic discrepancy between them as follow:

$$\mathcal{L}_{perc} = ||\phi_l(I_o) - \phi_l(I_p)||_2, \tag{7}$$

where ϕ_l are the activation after relu4_2 layer in the VGG-19 network.

Contextual-Loss: To encourage our network to preserve more details from source appearance image, we employ the loss proposed in [33] as follow:

$$\mathcal{L}_{cont} = \sum_l \zeta_l[-log(\frac{1}{n_l}\sum_i \max_j A^l(\phi_l(I_o), \phi_l(I_a)))], \tag{8}$$

where i and j index the feature map of layer ϕ_l that contains n_l features, and ζ_l controls the relative importance of different layers. A_l denotes the pairwise affinities between features. We use relu2_2 up to relu5_2 layers for ϕ_l.

Style-Loss: In order to obtain the more realistic output, we penalize the statistic error between high-level features as follow:

$$\mathcal{L}_{style} = \sum_{l} ||G_l(I_o) - G_l(I_a)||_2, \tag{9}$$

where G_l denotes the Gram matrix estimated from ϕ_l form relu2_2 to relu4_2.

Self-Loss: To fully utilize the data under self-supervision, we construct pseudo paired data by apply random geometry transformations to I_a to obtain its desired pose image I'_a. In this way, the output I_o should be the same as I'_o, we then penalize the loss as follow:

$$\mathcal{L}_{self} = \sum_{l} ||\phi_l(I_o) - \phi_l(I'_a)||_1, \tag{10}$$

where ϕ_l denotes the activation of layer form relu2_2 to relu5_2.

Regularization-Loss: In the matching block, since we align the image and mask from two domains, we here apply a $\mathcal{L}_1$ regularization to encourage them to be closer as follow:

$$\mathcal{L}_{reg} = ||f_a - f_p||_1. \tag{11}$$

Adversarial-Loss: To force the generator to generate realistic images, we deploy a discriminator like in [61] to discriminate the generated fake images. The adversarial objectives of D and G are respectively formulated as follow:

$$\begin{aligned} \mathcal{L}^D_{adv} &= -\mathbb{E}[(D(I_a))] - \mathbb{E}[-D(G(I_a, I_p))], \\ \mathcal{L}^G_{adv} &= -\mathbb{E}[D(G(I_a, I_p))]. \end{aligned} \tag{12}$$

Finally, we optimize the total loss as follow:

$$\begin{aligned} \mathcal{L}_{total} = \lambda_1\mathcal{L}_{geo} + \lambda_2\mathcal{L}_{tps} + \lambda_3\mathcal{L}_{perc} + \lambda_4\mathcal{L}_{cont} \\ + \lambda_5\mathcal{L}_{style} + \lambda_6\mathcal{L}_{self} + \lambda_7\mathcal{L}_{reg} + \lambda_8\mathcal{L}_{adv}, \end{aligned} \tag{13}$$

where $\lambda_1 \sim \lambda_8$ are hyper-parameters controlling the weights to balance the objectives.

4 Experiment

Implementation. We adopt Adam [18] with $\beta_1 = 0.1$, $\beta_2 = 0.999$ as the optimizer in our all experiments using PyTorch library. Our model is jointly trained for 200 epochs with input-size = 256×256. we set the learning rates to 0.0001 and 0.0004 respectively, for the generator and discriminator. We set $\alpha = 100$, $\lambda_1 = \lambda_2 = 1$, $\lambda_3 = 0.1$, λ_4, $\lambda_6 = 1$, $\lambda_5 = 0.01$, $\lambda_7, \lambda_8 = 10$. Let $C_{k_i S_j}$ denote a convolution layer with kernel size of i and a stride of j, followed by InstanceNorm2d Normalization [57] and ReLu activation function [36]. Let ResBlock denote the Residual Block structure proposed by [10], in which the BatchNorm2d Normalization is replaced by InstanceNorm2d Normalization [57]. Similar to [68], the

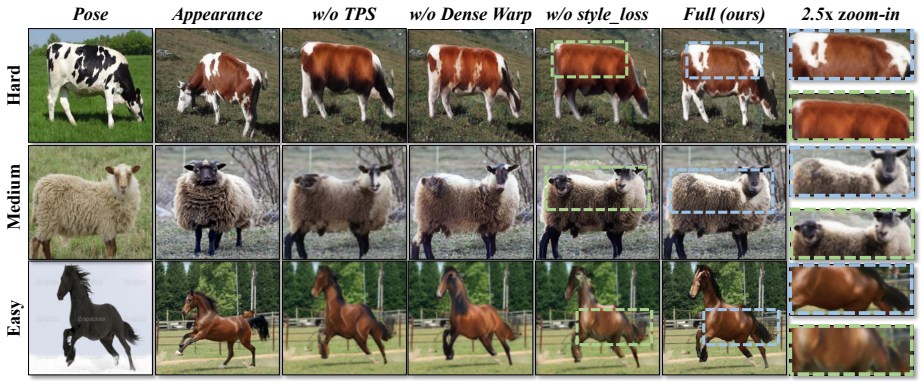

Fig. 3. Visual comparisons of our ablation methods.

Methods	*Human*		*Mammals*		*Birds*		*Cars*	
	mFID ↓	mSSIM ↑	mFID ↓	mSSIM ↑	mFID ↓	mSSIM ↑	mFID ↓	mSSIM ↑
w/o $\mathcal{L}_{perc}$	63.1	0.255	68.4	0.193	57.6	0.211	71.4	0.132
w/o $\mathcal{L}_{cont}$	52.8	0.406	58.5	0.394	49.2	0.301	62.2	0.267
w/o $\mathcal{L}_{self}$	40.6	0.601	36.6	0.521	30.7	0.528	43.1	0.452
w/o $\mathcal{L}_{reg}$	38.8	0.649	35.0	0.533	29.5	0.538	41.6	0.464
w/o Tps	38.9	0.645	35.0	0.547	29.3	0.527	41.6	0.459
w/o Dense Warp	38.7	0.651	34.8	0.554	28.9	0.539	41.3	0.464
w/o $\mathcal{L}_{style}$	39.1	0.635	35.2	0.523	29.8	0.517	41.9	0.448
TPS (TSC) ↔ Dense Warp (SS)	38.2	0.655	34.3	0.543	29.1	0.542	41.3	0.468
***Ours* (full)**	**37.6**	**0.676**	**33.9**	**0.576**	**28.3**	**0.571**	**40.8**	**0.491**

Table 1. Exploration of different components of our method. ↔ denotes the position change of **SS** and **TSC** blocks where they inject to.

two separate feature extractors F_A and F_B share the same structure but without sharing weight in the form of $\{C_{k_3S_1}, C_{k_4S_2}, C_{k_3S_1}, C_{k_4S_2}, C_{k_3S_1}, \text{ResBlock} * 3\}$, which will output two different features. Note that we perform different $C_{k_1S_1}$ and $C_{k_4S_4}$ operations to estimate the C_{dense} and C_{tps}, and thus these two matrices are in different shapes. Let L denote a linear function output m dimensions. As for the TPS warping regression network, we follow [59] and adopt the structure in the form of $\{C_{k_4S_2}, C_{k_4S_2}, C_{k_3S_1}, C_{k_3S_1}, L_{18}\}$. The foreground generator is in a encoder-decoder like network and the background generator network can be in arbitrary Unet [37] structure for reconstruction. More network architecture can be referred to supplementary material.

Datasets. We evaluate our method on several challenging datasets that contain large variations in terms of pose and category. Specifically, to illustrate that our framework can conduct pose transformation for general objects, we benchmark our method using four datasets: (i) *Human*: We perform training on the DeepFashion dataset provided by [26]. Note that we do not use the skeleton information and the ground-truth targeted pose images for training. (ii) *Mam-*

mals: We collect 5 classes of animals images including *horse*, *cow*, *sheep*, *dog* and *lion* from ImageNet [45] and WebDataset [48]. We then combine them to build the Mammals dataset. In total, it consists of ∼5k images, and we split them into training/testing set at the ratio of 8:2 on each subject separately. (iii) *Birds*: We use the Caltech-UCSD Birds-200-2011 [58] as our Birds dataset. We follow the setting as the original list for training and testing. (iv) *Cars*: We use VeRi [25] dataset as our Cars dataset which contains many categories with diverse poses, and we strictly follow the training and testing set as in the original paper.

Evaluation metrics. We use the mean Fréchet Inception Score (mFID) [11] and mean Structural Similarity (mSSIM) [63] to measure the distance between the distributions of transferred synthesized images and original real images. We also conduct a user study to compute user preference (UP) scores on the translation results. Specifically, given 30 images from each method randomly, we interviewed 1,00 participants and asked them to rate their favorite works. Note that the participants are unaware of the specific algorithm that produce the transferred images, and we finally report the proportion of the results they prefer.

4.1 Ablation Studies.

Table 1 reports all the results of our ablation experiments. Specifically, in four datasets, our full model outperforms others by different degrees. It shows that removing some kinds of blocks and loss functions, it will make the network learn less detailed appearance or spatial information. In addition, changing SS and TSC position can not yield better results. We conjecture that the encoder can retain the underlying appearance information, while the decoder is responsible for incorporating the high-semantic spatial transformation information. We further visualize some examples and make qualitative comparison, as shown in Fig 3. We show three cases including *cow*, *sheep* and *horse*, and we define them as hard, medium, easy examples according to their pose and appearance difficulties. *w/o* $\mathcal{L}_{style}$ and *w/o* Tps will cause the network to miss some detailed appearance information such as textures, style and colour, *i.e.*, the skin of cow, which makes the output results less realistic. Besides, although the visual performance of *w/o* Dense warp is close to the full model due to the powerful loss functions driven, it fails to achieve the satisfactory results in terms of mFID and mSSIM metrics.

4.2 Qualitative comparison.

Since general object pose transformation has not been extensively studied, therefore, we re-implement and compare with some existing generative models which can be applied to our task. Specifically, for *Mammals* and *Birds* datasets, we compare our method with CoCosNet [68], FUNIT [21] and COCO-FUNIT [46], where the former one aims to synthesize realistic images given the examplar images while the latter two focus on style-content translation. They both can be readily applied to our tasks, and we retrain the methods with the same training set as ours to keep the fairness. For *Human* dataset, we use the state-of-the-art

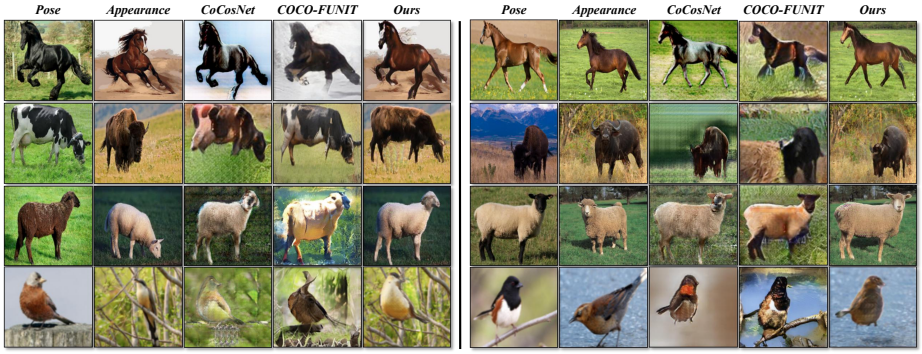

Fig. 4. Qualitative comparison of different methods on *Mammals* and *Birds* dataset.

Fig. 5. Qualitative comparison of different methods on *Human* and *Cars* datasets.

method Liquid-GAN [23] and Liquid++ [24] for comparisons. For *Cars* dataset, we compare our method with PAGM [29] and HoloGan [41], which can be applied to car view-synthesis. Note that all the re-implement results are produced by using the open-source code.

As shown in Fig 4, it shows that our model synthesizes more convincing results with well-preserved characteristics of objects. To be specific, Although CoCosNet [68] yields the transferred outputs, it fails to preserve the same appearance details from the source images. For instance, the output "*horse*" should be in brown body and with a little white on head rather than in the whole black body. Moreover, it can not fix the original background information, making the output images unsatisfactory. Likewise, FUNIT [21] and COCO-FUNIT [46] also fail to predict high-quality results since it transfers the style from one image to another globally, which will also break the background information. As for the foreground object characteristics, it cannot well deal with the issue of retaining the content from the source appearance images. Compared to these methods, our proposed method can successfully yield the output in the desired pose retaining the appearance of both objects and background due to the proposed sub-blocks and loss functions. Table 2 shows the detailed quantitative metrics comparison between these methods, among which, our method outperform them by a large margin and achieve the top user preference.

Methods	*Mammals*			*Birds*		
	mFID ↓	mSSIM ↑	UP ↑	mFID ↓	mSSIM ↑	UP ↑
FUNIT [21]	78.5	0.138	4%	80.4	0.182	3%
COCO-FUNIT [46]	70.7	0.141	7%	78.8	0.186	4%
CoCosNet [68]	81.6	0.156	6%	64.5	0.211	6%
Ours	**33.9**	**0.576**	**83%**	**28.3**	**0.571**	**87%**

Table 2. Quantitative comparisons of our method with other methods on *Mammals* and *Birds* dataset.

Methods	*Human*			*Cars*		
	mFID ↓	mSSIM ↑	UP ↑	mFID ↓	mSSIM ↑	UP ↑
Liquid [23]	44.6	0.559	20%	-	-	-
Liquid++ [24]	41.4	0.567	30%	-	-	-
HoloGan [41]	-	-	-	51.8	0.251	18%
PAGM [29]	-	-	-	46.7	0.284	28%
Ours	**37.6**	**0.676**	**50%**	**40.8**	**0.491**	**54%**

Table 3. Quantitative comparisons of our method with other methods on *Human* and *Cars* dataset.

For *Human* dataset, Fig 5 (left) shows that our method can produce more reasonable results. Note that our method is a generic general object pose transformation framework but not specially designed for human. In other words, we do not employ Face-Loss [22] and some keypoints or body mesh [27] information to train our network as in [23,24]. Therefore, the transferred output images we yield are acceptable and make it convenient to conduct human pose transformation without using auxiliary information. More quantitative comparisons can be seen in Table 3.

For *Cars* dataset, our method outperforms the recent state-of-the-art PAGM [29] and HoloGan [41] as shown in Table 3. More specifically, Fig 5 (right) shows us some examples that the previous methods fail to retain the color of the car (upper) or abortively transfer the view of the vehicle (bottom). As for our proposed framework, we successfully synthesize the new view of vehicle given the unpaired data, which illustrates the effectiveness of our approach.

4.3 Multi-object Pose Transformation.

To further illustrate the generality of our framework, we conduct experiments and visualize some multi-object examples. As shown in Fig 6 (left), given the same number of objects in appearance and desired pose images, we can transfer all the objects in the same pose, which broadens the usefulness of our framework. It's worth mentioning that we do not advocate transferring images with mismatched number of objects between appearance and pose images. This is

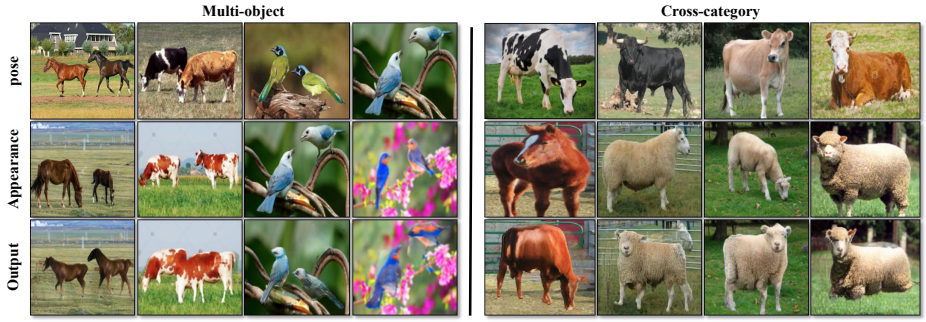

Fig. 6. Qualitative results of multi-object pose transformation and cross-category object pose transformation.

Network	CAM	Pseudo-Masks	Seg. Masks (val)	Seg. Masks (test)
IRNet [1]	48.3	65.9	63.5	64.8
+DA	$49.6_{+1.3}$	$66.8_{+0.9}$	$64.6_{+1.1}$	$65.8_{+1.0}$
SEAM [62]	55.4	63.4	64.5	65.7
+DA	$56.4_{+1.0}$	$64.4_{+1.1}$	$65.7_{+1.2}$	$66.5_{+0.8}$

Table 4. Different baselines performance with data augmentation by our framework in mIoU on PASCAL VOC dataset. **+DA** denotes conducting data augmentation.

because asymmetrical quantities will lead to ambiguity in object pose transformation.

4.4 Cross-category Pose Transformation.

Under some extreme circumstances, some species often have only a few images, and it is difficult to construct unpaired pairs. Take "*horse*" as an example, in real life, we may only observe the standing horses, and there is no other reference data. However, there are many other similar animals in nature, such as cow and sheep, which they all have limbs. Based on cross-category observations, we humans have the ability to imitate "*horse*" to an unknown posture based on references from other categories. With this in mind, we conduct experiments on cross-category pose transformation. As shown in Fig 6 (right), our proposed method can address this issue and produce reasonable images, which will not miss the original appearance details. This finding will encourage wider applications in the future.

4.5 Applications.

Data Augmentation. In the weakly supervised semantic segmentation (WSSS) task [1,56], it aims to leverage the class-activation-maps [52] to find out the ob-

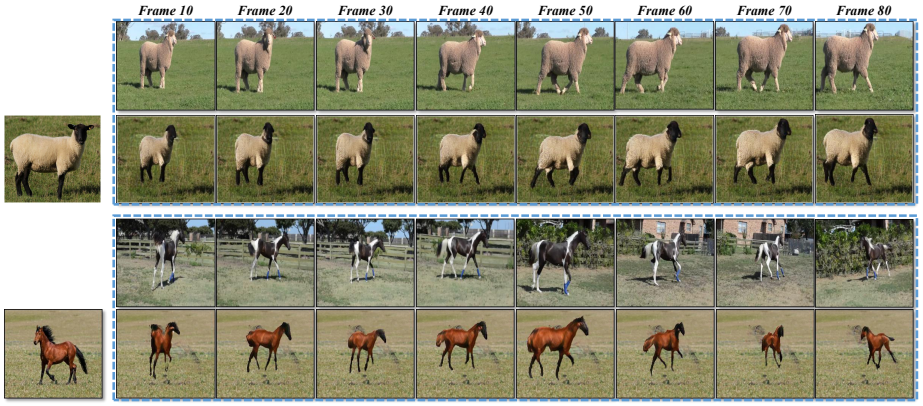

Fig. 7. Illustrative examples of objects video imitation. The reference pose videos are sampled from Got-10k [12] dataset. For more examples, please refer to supplemental material.

jects' potential regions and yield the pseudo masks with only class-level labels. In order to improve the class-activation-maps, object diversity is of great significance in weakly supervised semantic segmentation. In this setup, our goal is to expand the object images in different poses so as to provide more realistic images for training. We here choose IRNet [1] and SEAM [62] as baseline models to conduct experiments on PASCAL VOC dataset [8] to verify the quality of the images we produce. Note that we train our **UFO-PT** using PASCAL VOC dataset without extra data. Specifically, we produce more training images in different poses by our method, including "*horse*", "*cow*", "*sheep*", "*dog*", "*cat*", "*person*" and "*vehicle*". Table 4 shows that compared to the baseline model, our method can help both the baselines boost the performance, which demonstrates the practicality of our method.

Video Imitation. Moreover, we can apply our method to video imitation. As shown in Fig 7, given a static image providing appearance and a dynamic sequential video, we can yield an unseen video of that object. This intriguing study has a wide range of applications, which can synthesize more action videos of objects to reduce the burden of collecting data artificially.

4.6 Failure Case.

While our approach effectively addresses the general object pose transformation problem, it still has several failure modes. Fig 8 illustrates some failure cases generated by our method. When the body part of the image is hard to localize, the model generates unsatisfactory results. However, this is a common problem in deep learning, such as object occlusions and pose extraction of hard examples. We will try to alleviate this issue in future work to further improve our method.

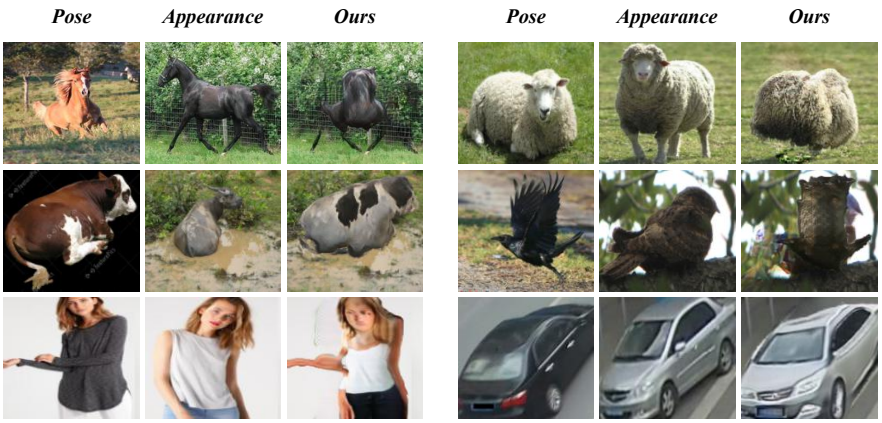

Fig. 8. Failure cases visualization.

5 More-In-Depth Discussion.

This paper pushes the frontier of the general object pose transformation that is beneficial to many applications in computer vision tasks. The final generation might be upper-bounded by the quality of the saliency mask, however, most of the previous works also adopt off-the-shelf methods like Densepose [9] or SMPL [27] to segment out the object for transformation. Segmenting out the foreground is not the main focus of our method. We take the early step to conduct general pose transformation, and we use the unsupervised saliency to highlight the foreground objects and conduct warping and generating images, which is acceptable.

6 Conclusion

Unlike the previous works that only focus on whether humans or some mammals in isolation, we introduce a unified framework for general object pose transformation with unpaired data. We propose to align and match two input images semantically using SS block and TSC block to inject spatial and detailed style information into the generating block. Experiments on different datasets show the superiority of our approach. Moreover, we can apply our framework to several applications such as data augmentation and video imitation, which can further show its practicality.

Acknowledgment. This work was supported by National Natural Science Foundation of China (NSFC) 61876208, Key-Area Research and Development Program of Guangdong Province 2018B010108002, and the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-RP-2018-003), the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (MOE-T2EP20220-0007) and Tier 1 (RG95/20).

References

1. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2209–2218 (2019) 12, 13
2. Balakrishnan, G., Zhao, A., Dalca, A.V., Durand, F., Guttag, J.: Synthesizing images of humans in unseen poses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8340–8348 (2018) 1, 3
3. Bansal, A., Sheikh, Y., Ramanan, D.: Shapes and context: In-the-wild image synthesis & manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2317–2326 (2019) 4
4. Cao, Y., Zhou, Z., Zhang, W., Yu, Y.: Unsupervised diverse colorization via generative adversarial networks. In: Joint European conference on machine learning and knowledge discovery in databases. pp. 151–166. Springer (2017) 1
5. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence **43**(1), 172–186 (2019) 2, 3
6. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5799–5809 (2021) 3
7. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2758–2766 (2015) 3
8. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015) 13
9. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7297–7306 (2018) 14
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 7
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. arXiv preprint arXiv:1706.08500 (2017) 9
12. Huang, L., Zhao, X., Huang, K.: Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 13
13. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017) 1
14. Ji, D., Kwon, J., McFarland, M., Savarese, S.: Deep view morphing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2155–2163 (2017) 3
15. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7122–7131 (2018) 2

16. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019) 4
17. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: International Conference on Machine Learning. pp. 1857–1865. PMLR (2017) 4
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 7
19. Kulkarni, T.D., Whitney, W., Kohli, P., Tenenbaum, J.B.: Deep convolutional inverse graphics network. arXiv preprint arXiv:1503.03167 (2015) 3
20. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017) 1
21. Liu, M.Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., Kautz, J.: Few-shot unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10551–10560 (2019) 4, 9, 10, 11
22. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220 (2017) 11
23. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5904–5913 (2019) 1, 3, 10, 11
24. Liu, W., Piao, Z., Tu, Z., Luo, W., Ma, L., Gao, S.: Liquid warping gan with attention: A unified framework for human image synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 10, 11
25. Liu, X., Liu, W., Mei, T., Ma, H.: Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. IEEE Transactions on Multimedia pp. 1–1 (2018) 9
26. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1096–1104 (2016) 8
27. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 1–16 (2015) 3, 11, 14
28. Lorenz, D., Bereska, L., Milbich, T., Ommer, B.: Unsupervised part-based disentangling of object shape and appearance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10955–10964 (2019) 4
29. Lv, K., Sheng, H., Xiong, Z., Li, W., Zheng, L.: Pose-based view synthesis for vehicles: A perspective aware method. IEEE Transactions on Image Processing **29**, 5163–5174 (2020) 4, 10, 11
30. Ma, L., Jia, X., Georgoulis, S., Tuytelaars, T., Van Gool, L.: Exemplar guided unsupervised image-to-image translation with semantic consistency. arXiv preprint arXiv:1805.11145 (2018) 4
31. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. arXiv preprint arXiv:1705.09368 (2017) 1, 3
32. Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B., Fritz, M.: Disentangled person image generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 99–108 (2018) 1, 3

33. Mechrez, R., Talmi, I., Zelnik-Manor, L.: The contextual loss for image transformation with non-aligned data. In: Proceedings of the European conference on computer vision (ECCV). pp. 768–783 (2018) 6
34. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014) 3
35. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4500–4509 (2018) 1
36. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Icml (2010) 7
37. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: Edgeconnect: Structure guided image inpainting using edge prediction. In: The IEEE International Conference on Computer Vision (ICCV) Workshops (Oct 2019) 8
38. Neverova, N., Guler, R.A., Kokkinos, I.: Dense pose transfer. In: Proceedings of the European conference on computer vision (ECCV). pp. 123–138 (2018) 3
39. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016) 4
40. Nguyen, D.T., Dax, M., Mummadi, C.K., Ngo, T.P.N., Nguyen, T.H.P., Lou, Z., Brox, T.: Deepusps: Deep robust unsupervised saliency prediction with self-supervision. arXiv preprint arXiv:1909.13055 (2019) 5
41. Nguyen-Phuoc, T., Li, C., Theis, L., Richardt, C., Yang, Y.L.: Hologan: Unsupervised learning of 3d representations from natural images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7588–7597 (2019) 3, 10, 11
42. Park, E., Yang, J., Yumer, E., Ceylan, D., Berg, A.C.: Transformation-grounded image generation network for novel 3d view synthesis. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 3500–3509 (2017) 3
43. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2337–2346 (2019) 2, 6
44. Rematas, K., Nguyen, C.H., Ritschel, T., Fritz, M., Tuytelaars, T.: Novel views of objects from a single image. IEEE transactions on pattern analysis and machine intelligence **39**(8), 1576–1590 (2016) 3
45. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y 9
46. Saito, K., Saenko, K., Liu, M.Y.: Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder. arXiv preprint arXiv:2007.07431 **2** (2020) 4, 9, 10, 11
47. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: Graf: Generative radiance fields for 3d-aware image synthesis. arXiv preprint arXiv:2007.02442 (2020) 3
48. Shen, T., Lin, G., Shen, C., Reid, I.: Bootstrapping the performance of webly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1363–1371 (2018) 9
49. Si, C., Wang, W., Wang, L., Tan, T.: Multistage adversarial losses for pose-based human image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 118–126 (2018) 3

50. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in Neural Information Processing Systems **32**, 7137–7147 (2019) 3
51. Siarohin, A., Sangineto, E., Lathuiliere, S., Sebe, N.: Deformable gans for pose-based human image generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3408–3416 (2018) 1, 3
52. Su, Y., Lin, G., Hao, Y., Cao, Y., Wang, W., Wu, Q.: Self-supervised object localization with joint graph partition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2289–2297 (2022) 12
53. Su, Y., Lin, G., Sun, R., Hao, Y., Wu, Q.: Modeling the uncertainty for self-supervised 3d skeleton action representation learning. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 769–778 (2021) 3
54. Su, Y., Lin, G., Wu, Q.: Self-supervised 3d skeleton action representation learning with motion consistency and continuity. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13328–13338 (2021) 3
55. Su, Y., Lin, G., Zhu, J., Wu, Q.: Human interaction learning on 3d skeleton point clouds for video violence recognition. In: European Conference on Computer Vision. pp. 74–90. Springer (2020) 2
56. Su, Y., Sun, R., Lin, G., Wu, Q.: Context decoupling augmentation for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7004–7014 (2021) 12
57. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016) 7
58. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011) 9
59. Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristic-preserving image-based virtual try-on network. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 589–604 (2018) 2, 5, 8
60. Wang, M., Yang, G.Y., Li, R., Liang, R.Z., Zhang, S.H., Hall, P.M., Hu, S.M.: Example-guided style-consistent image synthesis from semantic labeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1495–1504 (2019) 4
61. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018) 7
62. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12275–12284 (2020) 12, 13
63. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004) 9
64. Wu, B., Duan, H., Liu, Z., Sun, G.: Srpgan: perceptual generative adversarial network for single image super resolution. arXiv preprint arXiv:1712.05927 (2017) 1
65. Wu, W., Cao, K., Li, C., Qian, C., Loy, C.C.: Transgaga: Geometry-aware unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8012–8021 (2019) 4

66. Yang, F., Lin, G.: Ct-net: Complementary transfering network for garment transfer with arbitrary geometric changes. arXiv preprint arXiv:2105.05497 (2021) 2
67. Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., Luo, P.: Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7850–7859 (2020) 6
68. Zhang, P., Zhang, B., Chen, D., Yuan, L., Wen, F.: Cross-domain correspondence learning for exemplar-based image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5143–5153 (2020) 4, 5, 7, 9, 10, 11
69. Zhou, T., Tulsiani, S., Sun, W., Malik, J., Efros, A.A.: View synthesis by appearance flow. In: European conference on computer vision. pp. 286–301. Springer (2016) 3
70. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017) 4
71. Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: Progressive pose attention transfer for person image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2347–2356 (2019) 3