CT-Net: Complementary Transfering Network for Garment Transfer with Arbitrary Geometric Changes

Anonymous CVPR 2021 submission

Paper ID 4041

Abstract

Most existing virutal try-on methods attempt to transfer a stand-alone piece of clothing to an image of a person. While garment transfer shows greater potential in realistic applications with the goal of transfering outfits directly across different people images. However, transfering betwen images with heavy misalignments or severe occlusions still remains as a challenge. In this work, we propose Complementary Transfering Net (CT-Net) to adaptively model different levels of geomatric changes and transfer outfits between different people. In specific, CT-Net consists of three modules: i) a complementary warping module first estimates two complementary warpings to transfer the desired clothes and segmentations in different granularities. ii) The coarse transfered segmentations is further refined leveraging a mask refine module, which is then used to guide the generation. iii) A fuse module adaptively integrates all information to render the garment transfer results with well-preserved characteristics of clothes and identities of human. Extensive experiments conducted on DeepFashion dataset demonstrate that our model synthesizes high quality virtual try-on images and significantly outperforms all state-of-art methods both qualitatively and quantitatively.

1. Introduction

Most existing virtual try-on methods are based on simplifying assumptions: (i) Pure clothing images or 3D information are available as inputs; (ii) Pose changes are simple without heavy misalignments or severe occlusions of the clothes. We argue that these simplifying assumptions greatly limit the application scope of these methods in the realistic virtual try-on scenarios. To address this issue, we propose Complementary Transfering Network (CT-Net), a novel image-based garment transfer network that does not rely on pure clothing images or 3D information while capable to adaptively deal with different levels of geometric changes. As shown in Figure 1, given a target person image I^T and a model image I^M, without any restriction to the poses or shapes of I^T and I^M, our CT-Net synthesizes photo-realistic try-on results, in which the person in I^T wearing the clothes depicted in I^M with well-preserved details.

Figure 1. **Virtual try-on results generated by CT-Net. First row: model images. First column: target person images.** As shown above, CT-Net naturely transfers clothes across different people images with arbitrary poses or shapes and synthesizes photo-realistic try-on results with well-preserved characteristics of the desired clothes and identities of humans.

Despite various methods have been proposed to realize virtual try-on in different settings [13, 36, 41, 40, 29, 6, 11, 39], there is still a gap between these methods and the unlimited realistic scenarios. Some methods [10, 4, 28] in-

volve 3D information to deal with misalignments or occlusions, but they are greatly limited by expensive devices and high computational costs. [13, 36, 41, 40] rely on standalone clothing images, which are not easy to get timely online. Moreover, most of them attempt to model the geometric changes of the clothes utilizng a Thin Plate Spline (TPS) warping. However, limited by a small number of parameters, TPS warping is only capable to shape simple deformations, which makes their methods fail to deal with complex cases when heavy misalignments or severe occlusions occur. Garment transfer methods aim to transfer outfits across different people images. Although prior arts [29, 11, 39] has achieved considerable progress, none of them address the issue of heavy misalignments and severe occlusions.

We aim to fulfill this gap by proposing a novel garment transfer network, Complementary Transfering Network (CT-Net), which precisely transfers outfits across different people while tolerating heavy misalignments and severe occlusions. As shown in Figure 2, CT-Net has three sub-modules. First, a ComplementaryWarp Module(CWM) is introduced to warp the desired clothes and segmentation into target region. Specially, we estimate two complementary warpings with different levels of freedom: (i) Dense correspondence warping. (ii) Thin Plate Spline (TPS) warping. Dense correspondence warping has a high degree of freedom and is utilized to map pixels to be well-aligned with the target pose to guide the generation; While limited by a small number of parameters, TPS warping roughly transfers the desired clothes into target region with well-preserved details, which is utilized to further refine the synthesized result. Second, a MaskRefine Module(MRM) is introduced to predict transfered segmentation, in which the target person wearing the desired clothes. Opposited to prior works, which may suffer from the large misalignments between inputs [29, 11, 40], we adopt a transfer-refine strategy by first warping the original segmentation to be well-aligned with the target pose and then refine it to predict precise transfered segmentation. Our MaskRefine module not only works well in complex cases with heavy misalignments or severe occlusions, but also adds structural constraints to the learning of dense correspondence warping, encouraging the warping results to be more coherent with the source image. Third, a Fuse Module(FM) adaptively integrates all the information provided by previous modules to render the garment transfer results.

Extensive experiments conducted on DeepFashion dataset demonstrate the superiority of our method compared to the state-of-art methods. To summarize, our main contributions can be summarized as follows:

- We propose a novel image-based garment transfer network, which adaptively models different levels of geometric changes by estimating two complementary warpings and synthesizes photo-realistic garment transfer results with well-preserved details of the clothes and human parts.
- We introduce a novel ComplementaryWarp module to estimate two precise complementary warpings simultaneously.
- A novel MaskRefine Module is proposed to predict precise transferred segmentation, which also adds structural constraints to the training of warping module, forcing the warping results to be more coherent with target pose.
- Evaluated on DeepFashion [23] dataset, CW-Net outperforms all the state-of-art methods by a large margin and synthesizes photo-realistic images with well-preserved clothing and body details.

2. Related Work

Generative Adversarial Networks. Generative Adversarial Networks (GANs) [9] have been demonstrated very effective in generating fake images, which are indisguishable from the real ones in the original dataset. To further control the generation results, conditional GAN (cGAN) [26] inputs extra information to guide the generation, which promotes the development of many applicaitons, such as image editing [5, 43, 19, 15] and image synthesis [16, 44, 37, 27, 1, 12]. Specially, Isola *et al*. [16] proposed an image-to-image translation network to transfer images from one domain to another (*e.g*. segmentation to photos, edges to photos), which shows convincing ability to handle cross-domain relationships. Similarly, we also employ a cGAN to synthesize photo-realistic virtual try-on results conditioned on desired clothing and target pose representations.

Pose-guided Human Image Generation. Ma *et al*. [24] made an early attempt to generate human images conditioned on pose utilizing a two-stage network. They first generate a coarse image according to the target pose and then refine it with adversarial loss. Esser *et al*. [7] proposed a conditional U-Net to disentangle the pose and appearence. However, their model still suffers from misalignments between source and target pose, result in unsatisfying generation when large spatial transformation occurs. More recent methods propose to solve this problem utilizing warp-based methods. Siarohin *et al*. [33] introduces deformable skip connections to spatially transform the features. Zhu *et al*. [45] employs a sequence of pose-attentional transfer blocks to progressively deal with large pose discrepancies. [11, 30, 21] estimate appearence flow to transfer the pixels or features to be aligned with the target view to facilitate the generation. Zhang *et al*. [42] for the first time introduces cross-domain semantic matching, which precisely learns a dense mapping between cross-domain inputs. Inspired by [42], we also estimate a dense semantic warping. However, we focus on the exact problem of garment transfer and learn

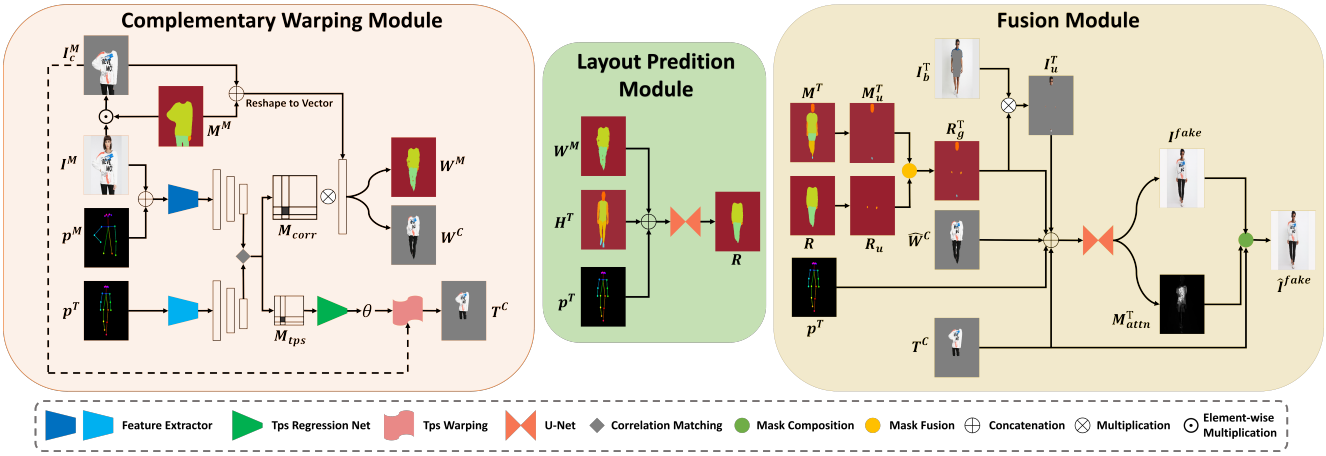

Figure 2. The overall architecture of CT-Net, which consists of three modules. (i) ComplementaryWarp Module estimates two complementary warpings to transfer the desired clothes in different granularities. (ii) MaskRefine Module predicts precise transfered segmentations based on coasre transfered result W^M. (iii) Fuse Module adaptively integrates all information to render photo-realistic garment transfer results.

a semantic maping between pose representations. Benefited from the joint training of all the modules, we achieve more precise warping results on the clothing items.

Virtual Try-on. Virtual Try-on has been paid more and more attention as the fast development of e-commerce. Many conventional virtual try-on methods rely on 3D information, such as pre-defined 3D clothing template [10], 3D body model [4] or depth information captured by 3D sensors [28]. Along with the advances of deep neural networks, more recent works attempt to synthesize try-on results based on 2D images. Han *et al*. [13] proposed VITON to transfer clothes in a stand-alone pure clothing image to the target person. Variations [36, 41, 40] are proposed to improve VITON with clearer clothing details and well-preserved human identities. However, all of them rely on simple assumptions that the clean clothing images are available and geometric changes are simple enough to be modeled by a thin plate spline warping (TPS) with a small number of parameters (*e.g*. 6 for affine and $2\times5\times5$ for TPS as in [36]). These assumptions greatly hinder the application of these methods in the realistic virtual try-on scenarios. Wu *et al*. [39] proposed to use densepose [2] descriptor to warp the desired clothes onto the target person. But mapping estmated by densepose descriptor can be very sparse when there are large occlusions, leading to unconvincing synthesized results. Methods mentioned above only focus on the transfer of upper clothes. SwapNet [29] employs a two-stage network to transfer the entire outfits of the inputing images. To deal with the misalignments of features, they adopt ROI pooling and encode each clothing regions into high-dimensional features before feeding them into the generator. However, the encoded features are inadequate to preserve the details of the local textures, which leads to blurry synthesized results. Different from these methods, we adaptively combine two complementary warpings to deal with different levels of deformations and synthesize photo-realistic images with well-preserved characteristics of clothes and identities of humans.

3. Complementary Transfering Network

Given the image I^M depicting model wearing desired clothes, the image I^P depicting target person, assuming clothes, poses and shapes of I^M and I^P can be arbitrary, our goal is to synthesize high quality try-on results with well-preserved textures. To achieve our goal, we adopt a warp-refine strategy, presenting Complementary Transfering Network (CT-Net). As shown in Figure 2, CT-Net consists of three modules. First, we introduce a ComplementaryWarp Module (CWM), in which a dense correspondence warping and a Thin Plate Spline warping are estimated. Dense correspondence warping is utilized to warp the clothes and segmentations of I^M to be aligned with target poses to guide the generation, and the synthesized result will be further refined leveraging the complementary warping result from TPS (Section 3.1). Second, we introduce a MaskRefine Module (MRM), in which we refine the transfered mask with extra clothing-agnostic pose representations (Section 3.2). The third module is a Fuse Module (FM), which adaptively integrates all the information to render photo-realistic garment transfer results (Section 3.3).

3.1. ComplementaryWarp Module

To synthesize garment transfer results, one of the main challenges is to combine the clothes of the model with the target person's pose. Most of the prior works feed the target

pose and model clothes into conditional generative adversarial netwrok (cGAN). However, in cases with heavy misalignments or severe occlusions, they can only retain the global style and fail to preserve local details. A good practice to overcome this issue is to utilize warping methods to align the clothes with the target pose first before feeding them into generation. Similarly, we propose ComplementaryWarp Module, which simultaneously estimates two complementary warpings, to transfer the information in two different granularities to facilitate the generation.

As shown in Figure 2, we first employ two seperate feature extractors to extract high-level features. Then we match the features to calculate correspondence matrixs, $\mathcal{M}_{corr}$ and $\mathcal{M}_{tps}$, which are used to estimate dense correspondence warping and TPS warping. Given a model image I^M, the corresponding clothes I_c^M and the segmentation M^M, dense correspondence warping is utilized to transfer I_c^M and M^M in finer granularity to get W^C and W^M, which are well-aligned to target pose. TPS warping transfers I_c^M in a coarse granularity to get T^C, which is roughly aligned with target region.

Correspondence Matrix. We adopt the state-of-art pose estimation method [3] to estimate pose segmentations of the target person and the model image. And we further convert the sparse pose segmentations into dense by replacing each zero pixel with its distance to the joints, denoted as p^T, p^M. Different from [42], which aims to learn a correspondence matrix from multi-modal inputs, we simplify the task by directly inputing pose representations in both sides. To be specific, let $\mathcal{F}_A$, $\mathcal{F}_B$ denote the seperate feature extractors, we first extract high-level features $m_f \in \mathbb{R}^{H\times W\times C}$ and $t_f \in \mathbb{R}^{H\times W\times C}$ as follows:

$$m_f = \mathcal{F}_A(I^M, P^M) \quad (1)$$

$$t_f = \mathcal{F}_B(P^T) \quad (2)$$

To estimate the correspondence matrixs, we first aggregate the features into different scales with different kernels, which is illustrated in Figure 3. Specificaly, we use kernel of size 3, wiht stride 1 and padding size 1 to estimate the correspondence matrix $\mathcal{M}_{corr} \in \mathbb{R}^{HW\times HW}$ for dense correspondence warping and we select kernel of size 4, with stride 4 and padding size 0 to estimate $\mathcal{M}_{tps} \in \mathbb{R}^{HW/16\times HW/16}$ for TPS warping.

We employ the same correspondence layer as [42] to match the aggregated features m'_f and t'_f , which can be formulated as:

$$\mathcal{M}(i,j) = \frac{(m'_f(i)^T - u_m)(t'_f(j) - u_t)}{\|m'_f(i) - u_m\|\|t'_f(j) - u_t\|} \quad (3)$$

where u_m and u_t represent mean vectors.

Dense Correspondence Warping. We calculate the weighted average to estimate the dense correspondence warping [42]:

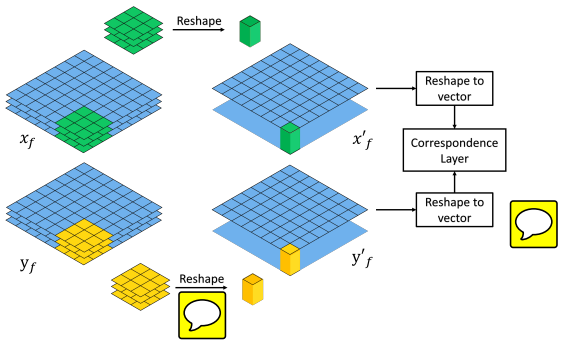

Figure 3. Illustration of the correspondence matching process with different kernels. This figure shows an example of kernel with size 3.

$$\mathcal{W}^X(u) = \sum_v softmax(\alpha \mathcal{M}_{corr}(u,v)) \cdot Y^X(v) \quad (4)$$

where α is a hyper-parameter controling the sharpness of the softmax. We set it as 100 here. Dense correspondence warping learns a dense mapping between two images with a high degree of freedom, which is capable to handle large geometric changes. We utilize it to transfer the clothes and segmentions of the model image to be well-aligned with the target person's pose and shape, which provides important guides for the generator to synthesize high-quality try-on results. However, learning a perfect dense semantic matching is hard, some important textures may get lost in the warping. To deal with this issue, we involve TPS warping, which has very limited parameters but is capable to transfer with little information loss, to refine the final result.

TPS Warping. We estimate the TPS warping from $\mathcal{M}_{tps}$. As shown in Figure 2, we first employ a regression net [6, 40] to predict the corresponding control points and then calculate the parameters θ.

Given the model clothes I_c^M, we transfer it with deformation shaped by TPS to be roughly aligned with the target person I^T. For training, we adpot the second-order constraint from Yang *et al.* [40] to restrict the TPS warping from generating unnatural deformations or mess textures, which can be formulated as:

$$\mathcal{L}_2 = \sum_{p\in P} \lambda_r |\|pp_0\|_2 + \|pp_1\|_2| + |\|pp_2\|_2 - \|pp_3\|_2| + \lambda_s(|S(p,p_0) - S(p,p_1)| + |S(p,p_2) - S(p,p_3)|) \quad (5)$$

Where λ_r and λ_s are hyper-parameters; $S(p,p_i) = \frac{y_i - y}{x_i - x} (i = 0,1,2,3)$ is the slope between two points. The total loss can be represented as $\mathcal{L}_{tps}$:

$$\mathcal{L}_{tps} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2 \quad (6)$$

where $\mathcal{L}_1 = \|I_c^T - T^C\|_1$; λ_1 and λ_2 are the weights for two loss terms. Both of them are set to 10, respectively, in our experiments.

3.2. MaskRefine Module

We propose MaskRefine Module to generate a precise transfered segmentation, in which the target person wearing the desired clothes depicted in the model image. Prior works mostly feed the pose representations of the target person and the model clothes into generative adversarial networks (GANs) directly to generate the transfered segmentations. However, these methods greatly suffer from the misalignments between the target pose and desired clothes thus fail to predict convincing results while facing with large geometric changes. Opposited to these methods, we explore a new warp-refine strategy. As shown in Figure 2, we first warp the original segmentation in CWM to make it well-aligned with the target person's pose, then we utilize a U-net strucutre to refine it with extra clothing-agnostic representations. In this way, we eliminate the misalignments between two inputs and make it much easier for our network to predict precise transfered layouts.

We get the clothing segmentations utilizing the state-of-art human parsing network [22] and apply the segmentation of densepose descriptor [2] as our extra clothing-agnostic representations H^T. For trainig this module, we adopt the pixel-level cross-entropy loss, denoted as $\mathcal{L}_{mask}$. Since the head and shoes are not in the transfer list of our task, we remove these areas in the segmentation before feeding it into our network. Through joint training, MaskRefine module also adds structural constraints to the training of CWM, forcing warping results to be more coherent with target pose and have clearer boundries.

3.3. Fuse Module

Fuse Module is proposed to adaptively integrate all the information provided by previous stages to render garment transfer results. As shown in Figure 2, we first fuse the head and shoes masks in the original segmentation with the refined arm and leg masks to form the guide segmentation R_g^T. Leveraging R_g^T, we extract the non-target body parts from the masked body region I_u^P. We then adopt conditional generative network (cGAN) to integrate the untarget body parts I_u^P, the guided mask R_g^T, warping results $\widehat{W}^C$ and T^C and the pose segmentations of the target person p^T to render the try-on result I^{fake}. Note $\widehat{W}^C = W^C \odot R_c$, where R_c denotes the refined clothing mask. By feeding the non-target body parts, the generator learns to preserve all the details in the non-target region and synthesize results with distinct human identities [40]. To further refine the generation result, we simultaneously predict a composition mask M_{attn}^T utilizing the generator. The generated result I^{fake} is composited with the warping result of TPS

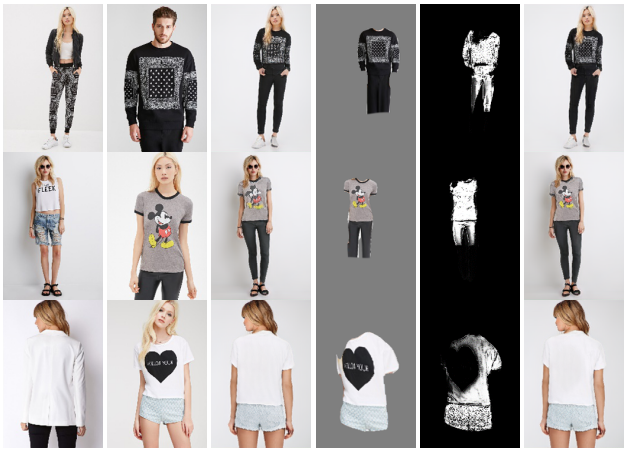

Figure 4. Examples of our attention mechanism. From left to right: target Person I^T, Model Image I^M, Generation Result I^{fake}, TPS warped Result T^C, Attention Mask M_{attn}, Refined Result $\hat{I}^{fake}$.

to synthesize the final garment transfer result $\hat{I}^{fake}$:

$$\hat{I}^{fake} = T^C \odot M_{attn}^T + I^{fake} \odot (1 - M_{attn}^T) \quad (7)$$

Leveraging the attention mask, we adaptively deal with different cases by combining the advantages of the two warping methods. For example, shown by the third row in Figure 4, when there are heavy misalignments or severe occlusions, the final generation result may rely more on the warping result from dense correspondence warping, ignoring the large logos or unreasonable textures; When the geometric change is simple and can be shaped by the TPS warping, as the first two rows in Figure 4, the attention mask may select more regions on the warping result from TPS to refine the generation results. In this way, we adaptively deal with different levels of geometric changes by combing the complementary warping results and expand the application scope of our model to wilder scenarios.

3.4. Loss Functions

To encourage the training of different modules benefit each other, we train our model in a joint style. We combine several different losses to produce high-quality try-on results, which will be introduced in details in the following sections:

Perceptual Loss. Based on differences between high-level features, perceptual loss has been proved effciently in the image generation tasks [18]. To pose perceptual constraints on the synthesized results, we adopt a pre-trained VGG network [34] to extract multi-level features ϕ_j to com-

pute the perceptual loss as:

$$\mathcal{L}_{perceptual}(G) = \sum_{j=1}^{N} \lambda_j \|\phi_j(\hat{x}_t) - \phi_j(x))\|_2 \quad (8)$$

Style Loss. We further adopt the style loss [8] to penalize the statistic error between high-level features, which can be formulated as:

$$\mathcal{L}_{style} = \sum_{j=1}^{N} \|G_j^{\phi}(\hat{x}_t) - G_j^{\phi}(x_t)\|_2 \quad (9)$$

where G_j^{ϕ} denotes the Gram matrix estimated from ϕ_j.

Contextual Loss. To encourage our network to preserve more details from the model clothes I_c^M, we employ the contextual loss proposed in [25], which can be formulated as:

$$\mathcal{L}_{contextual} = \sum_{l=1} \lambda_l \left[-\log \left(\frac{1}{n_l} \sum_i \max_j A^l(\phi_i^l(\hat{x}_B), \phi_j^l(y_B)) \right) \right], \quad (10)$$

where A^l denotes the pairwise affinities between features.

Adversarial Loss. To force the generator to learn the real distributions of the dataset and generate realistic human images, we deploy a discriminator given in [37] to discriminate the generated fake images from the real samples in the dataset. The loss can be formulated as:

$$\mathcal{L}_{adv}^{D} = \mathbb{E}_{x,y}[\log(\mathcal{D}(x,y))] + \mathbb{E}_x[\log(1 - \mathcal{D}(y, \mathcal{G}(x)))]$$
$$\mathcal{L}_{adv}^{G} = -\mathbb{E}[\mathcal{D}(\mathcal{G}(x,y))] \quad (11)$$

where x represents the input and y is the ground-truth.

Objective Fuction. Besides losses above, we apply a L1 regularization $\mathcal{L}_{reg} = \|1 - M\|_1$ on M_{attn} to encourage the generator to learn more from the TPS warped clothes. We also take L1 loss to stabilze our training process, which can be defined as $\mathcal{L}_{L1}(G) = \|\hat{x} - x\|_1$. Our objective function is a weighted sum of above terms:

$$\mathcal{L}(G) = \alpha_1\mathcal{L}_1 + \alpha_2\mathcal{L}_{tps} + \alpha_3\mathcal{L}_{mask} + \alpha_4\mathcal{L}_{perceptual} + \alpha_5\mathcal{L}_{style} + \alpha_6\mathcal{L}_{contextual} + \alpha_7\mathcal{L}_{adv} + \alpha_8\mathcal{L}_{reg} \quad (12)$$

where $\alpha_i, (i = 1, \ldots, 8)$ are hyper-parameters controling the weights of each loss.

4. Experiments

4.1. Dataset

We evaluate our model on the In-shop Clothes Retrieval Benchmark of DeepFashion dataset [23], which contains 52,712 fashion images of resolution 256 × 256. For training, we select 37,836 pairs of images depicting the same person wearing the same outfit with different poses. At test stage, we select 4,932 pairs of images which are not overlaped with the training set. As the realistic virtual try-on scenarios, each testing pair contains two different people with different clothes and poses.

4.2. Implementation Details

Architecture. Let Ck denote a Convolution layer with kernel size of 4, a stride 2, and k filters, followed by InstanceNorm2d Normalization [35] and ReLu activation function. Let Rk denotes the a Convolution layer with kernel size of 3, a stride 1, and k filters. Let Lk denotes a Linear function output k dimension. Let ResBlock denotes the original Residual Block proposed in [14], in which the BatchNorm2d Normalization is replaced by InstanceNorm2d Normalizaiton and the filters of the convolution layers are 256. The two seperate feature extractors $\mathcal{F}$ in CWM share the same structure: R64, C128, R256, C256, ResBlock * 6. We adopt the same regression net as [6] for TPS warping, which consists of C512, C256, C128, C64, L32. Structure of generators in MRM and FM are the same as U-Net [31] except the normalization is replaced by InstanceNorm2d Normalizaiton. All the discriminators are from pix2pixeHD [37].

Implementation. We adopt Adam [20] with $\beta_1 = 0.5, \beta_2 = 0.999$ as the optimizer in our all experiments. Our model is jointly trained for 100 epoches. Learning rate is fixed at 0.0002 for the first 50 epoches and then decays to zero linearly in the remaining steps. To balance the scales of losses in Eqn. 12, we set $\alpha_{1,2,3,7,8} = 10$ and $\alpha_{4,5,6} = 1$.

4.3. Baselines

ACGPN. ACGPN is a state-of-art virtual try-on network proposed by Yang *et al.* [40], which aims to transfer a standalone clothing image onto a reference person. In comparsion to previous methods [13, 36], ACGPN first predicts the transfered clothing segmentation in a two-stage strategy, then estimates the TPS warping by feedinig the transfered and orginal clothing segmentation into a STN [17]. ACGPN shows state-of-art performance on VITON [13] dataset with naturally transfered clothes and well-preserved untarget body parts.

CoCosNet. CoCosNet stands for the cross-domain correspondence network proposed by Zhang *et al.* [42], aiming to synthesize realistic image given the examplar images. Different from other methods, CoCosNet first align images in different domain by establishing dense correspondence matching and then use the aligned examplar image as guide to synthesize photo-realistic images, which shows state-of-art performance in pose-guided image generation task. However, CoCosnet lacks of the ability to generate

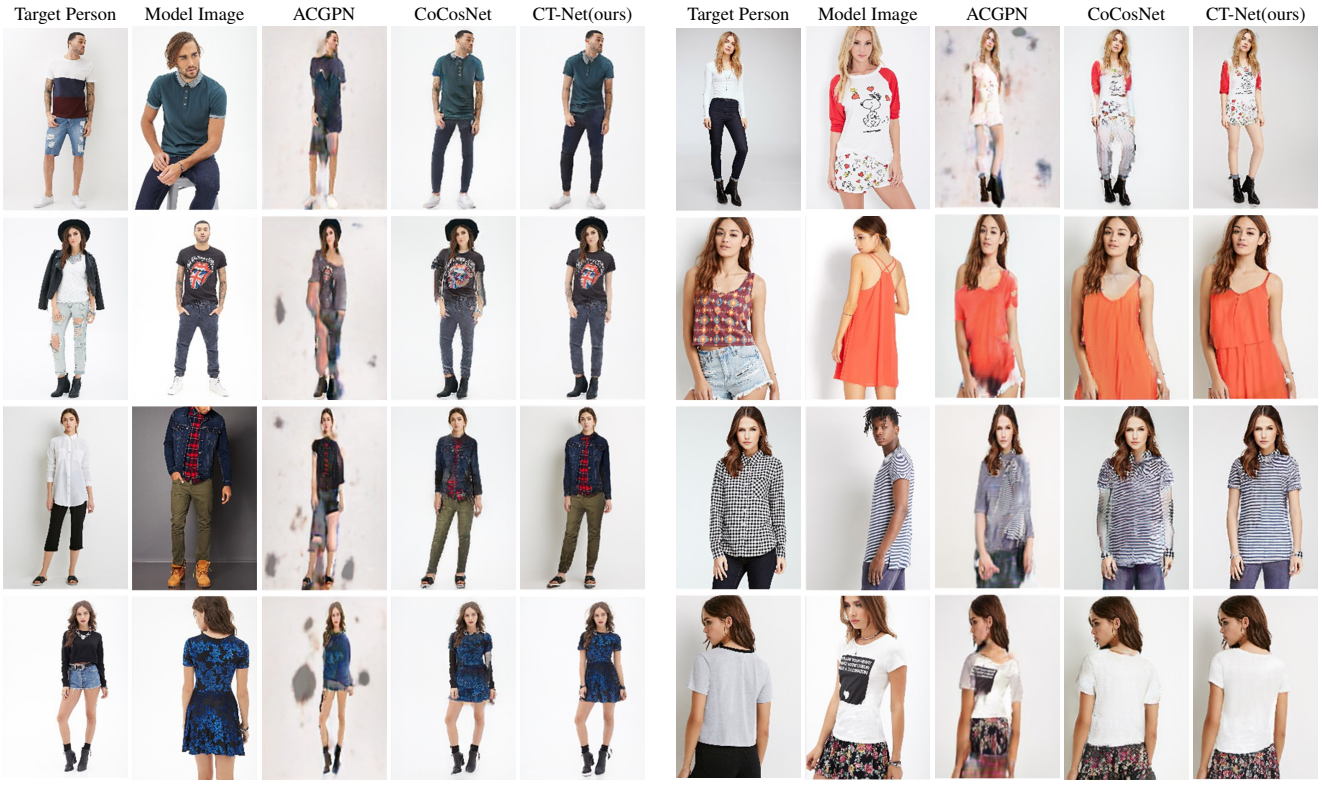

Figure 5. Qualitative comparisons of our method with other methods].

virtual try-on results. To adapt it into our task, we directly use the cross-domain correspondence warping to transfer the semantic segmentations and replace the inputs for the translation network in CoCosNet to be the same as ours in the FuseModule.

To keep the fairness of our experiments, We retrain all aforementioned methods on DeepFashion [23] dataset with the same training set as ours.

4.4. Qualitative Results

Figure 5 shows the qualitative comparisons of ACGPN, CoCosNet and our model, which shows that our model synthesizes more convincing results with well-preserved characteristics of clothes and identities of human. Although ACGPN [40] predicts transfered segmentation to stablize the estimation of TPS warping, it overlooks the misalignments between inputing segmentations. As shown in Figure 5, ACGPN fails to estimate precise transfered segmentations and TPS warping in cases with large deformations, leading to unsatisfying results with messy clothes, incorrect body parts and many noises. Compared to AGCPN, the CoCosNet [42] baseline addresses the misalignments by employing the cross-domain correspondence module to first align the input images. However, owing to the high degree of freedom, learning a perfect dense cross-domain correspondence warping is nearly impossible. Information loss occurs in the warping process, result in blurry warped clothes and coarse transfered segmentations. Without further refinement, visual artifacts such as clutted textures and blurry boundries can be observed in the synthesized results of CoCosNet. Different from ACGPN and CoCosNet, we propose a ComplementaryWarping Module by combing the advantages of dense correspondence warping with thin plate spline warping, which adaptively deals with different levels of geometric changes. As shown in the first row (right) and second row (left) of the Figure 5, logos on the model clothes are lost in the synthesized results of CoCosNet, while benefited from complementary warping, CT-Net preserves them clearly. In the last row (right), since ACGPN is unaware of the transformation of view points, it wrongly preserves the logo in the back view of the clothes, while CT-Net adaptively drop the logo and synthesizes reasonable back-view image.

In Figure 6, we visualize warping results from different methods to make further comparisions. Benefited from the joint training of all modules, CT-Net shows superior performance in estimating both the dense correspondence warping and the Thin Plate Spline warping more precisely.

Figure 6. Visual comparisons of warping results. Corr(CWM) represents the warping results of dense correspondence warping estimated in ComplementaryWarp Module.

Methods	Warp-SSIM	Mask-SSIM	H-SSIM	IS
ACGPN[40]	0.744	0.757	0.813	3.311
CoCosNet[42]	0.815	0.835	0.821	3.542
w/o PR	0.857	0.913	0.919	3.465
w/o MRM	0.836	0.917	0.923	3.452
w/o TPS	0.875	0.923	**0.937**	**3.497**
Name(ours)	**0.883**	**0.928**	0.930	3.479

Table 1. Quantitative comparisons of our method with other methods

4.5. Quantitative Results

We adopt Structural Similarity (SSIM) [38] to measure the similarity between generated images and the real ones and adopt Inception Score (IS) [32] to evaluate the quality of our synthesized images. To compute SSIM, we take the same person wearing the same clothes but in different poses as ground truth to compare with our synthesized images. Specificly, inspired by [33, 11], we compute SSIM in three different scales: (i) To isolate the influence from the background, we compute SSIM for human pixels (H-SSIM). (ii) To evaluate the accuracy of reconstructed clothes, we compute SSIM for clothing area of the synthesized results (Mask-SSIM). (iii) To compare the warping accuracy of different methods, we compute SSIM for warped clothes (Warp-SSIM). For our model, we use the warping result of dense correspndence warping to calculate Warp-SSIM.

Table 1 reports the quantitative results of our method and baselines. Higher scores are better. As summarized in table 1, our method outperforms all baselines by a significant margin. Specially, our model greatly improves the warping accuracy with 0.063 higher Warp-SSIM scores compared to CoCosNet. Moreover, we also achieve higher scores in terms of H-SSIM, Mask-SSIM and IS, which indicates that our method synthesizes more realistic images with well-preserved details.

Figure 7. Visual comparisons of our ablation methods.

4.6. Ablation Study

We conduct ablation experiments to explore the effectiveness of the main components in our model. In particular, *w/o* PR denotes removing the pose representation P^M inputing to the ComplementaryWarp Module. *w/o* MRM denotes removing MaskRefine Module. *w/o* TPS denotes removing the estimation of TPS warping.

Table 1 reports all the results of our ablation experiments. In specific, our full model outperforms *w/o* PR and *w/o* MRM by a margin in Warp-SSIM, which indicates that feeding p^M in the other side and adding structural constraints utilizing MaskRefine Module significantly facilitate the estimation of dense correspondence warping. We also find that our full model and *w/o* TPS have similar scores in all metrics, since SSIM only roughly measures the local similarity and IS only captures the realism of the images.

To further demonstrate the superiority of our full model, we visualize some examples to make qualitative comparsions in Figure 7. Since *w/o* PR and *w/o* MRM can not estimate the warping precisely, artifacts such as incorrect clothing shape (first row) and blurry boundries (third row) can be observed. Although *w/o* TPS achieves the best scores in terms of H-SSIM and IS, visual results show that our full model synthesizes more photo-realistic images with clearer and more complete clothing patterns.

5. Conclusion

We propose Complementary Transfering Net (CTNet) for garment transfer with aribitrary geometric changes. In particular, our model addresses the issue of transfering outfits with high misalignments or severe occlusions, while preserving the characteristics of clothes and the identity of humans well. We introduce three novel modules: i) We

propose a Complementary Warping Module (CWM), which estimates two complementary warpings to transfer the desired clothes and segmentations in different granularities. ii) A Mask Refine Module (MRM) is employed to refine the coarse transfered segmentations. iii) A Fuse module is proposed to adaptively integrate all the information to synthesize photo-realistic garment transfer results. Experiment results demonstrate that our model significantly outperforms existing state-of-art methods both qualitatively and quantitatively.

References

[1] Badour AlBahar and Jia-Bin Huang. Guided image-to-image translation with bi-directional feature transformation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9016–9025, 2019. 2

[2] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 3, 5

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 4

[4] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3d pose estimation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 479–488. IEEE, 2016. 1, 3

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

[6] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1161–1170, 2019. 1, 4, 6

[7] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018. 2

[8] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in neural information processing systems*, pages 262–270, 2015. 6

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[10] Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. Drape: Dressing any person. *ACM Transactions on Graphics (TOG)*, 31(4):1–10, 2012. 1, 3

[11] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10471–10480, 2019. 1, 2, 8

[12] Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R Scott, and Larry S Davis. Finet: Compatible and diverse fashion image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4481–4491, 2019. 2

[13] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018. 1, 2, 3, 6

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[15] Wei-Lin Hsiao, Isay Katsman, Chao-Yuan Wu, Devi Parikh, and Kristen Grauman. Fashion++: Minimal edits for outfit improvement. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5047–5056, 2019. 2

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[17] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 6

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[21] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019. 2

[22] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):871–885, 2018. 5

[23] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 2, 6, 7

[24] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image genera-

tion. In *Advances in neural information processing systems*, pages 406–416, 2017. 2

[25] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 768–783, 2018. 6

[26] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[27] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 2

[28] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (TOG)*, 36(4):1–15, 2017. 1, 3

[29] Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. Swapnet: Image based garment transfer. In *European Conference on Computer Vision*, pages 679–695. Springer, 2018. 1, 2, 3

[30] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2020. 2

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 6

[32] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016. 8

[33] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3416, 2018. 2, 8

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[35] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 6

[36] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 589–604, 2018. 1, 2, 3, 6

[37] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 2, 6

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 8

[39] Zhonghua Wu, Guosheng Lin, Qingyi Tao, and Jianfei Cai. M2e-try on net: Fashion from model to everyone. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 293–301, 2019. 1, 2, 3

[40] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. 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*, pages 7850–7859, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[41] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10511–10520, 2019. 1, 2, 3

[42] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 2, 4, 6, 7, 8

[43] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. *arXiv preprint arXiv:2004.00049*, 2020. 2

[44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2

[45] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019. 2