# Full-Range Virtual Try-On with Recurrent Tri-Level Transform

Han Yang[1,2] Xinrui Yu[3] Ziwei Liu[4] ✉
[1]ZMO.AI [2]ETH Zurich
[3]Harbin Institute of Technology, Shenzhen [4]S-Lab, Nanyang Technological University
hanyang@ethz.ch, stuyxr@stu.hit.edu.cn, ziwei.liu@ntu.edu.sg


Figure 1. Standard clothes such as T-shirts and long-sleeve jackets are well-analyzed in virtual try-on methods, while non-standard clothes with irregular design and patterns are scarcely reported. Given a target clothing image and the reference person, our method can synthesize photo-realistic results with accurate clothing shape, regardless of the clothing type. Hard samples including off-shoulder clothes and the word-shoulder clothes are covered in our framework, while the baseline methods fail to generalize.

## Abstract

*Virtual try-on aims to transfer a target clothing image onto a reference person. Though great progress has been achieved, the functioning zone of existing works is still limited to* ***standard clothes*** (e.g., *plain shirt without complex laces or ripped effect), while the vast complexity and variety of* ***non-standard clothes*** (e.g., *off-shoulder shirt, word-shoulder dress) are largely ignored.*

*In this work, we propose a principled framework,* ***Recurrent Tri-Level Transform (RT-VTON)****, that performs full-range virtual try-on on both standard and non-standard clothes. We have two key insights towards the framework design:* ***1) Semantics transfer*** *requires a gradual feature transform on three different levels of clothing representations, namely clothes code, pose code and parsing code.* ***2) Geometry transfer*** *requires a regularized image deformation between rigidity and flexibility. Firstly, we predict the semantics of the "after-try-on" person by recurrently refining the tri-level feature codes using local gated attention and non-local correspondence learning. Next, we design a semi-rigid deformation to align the clothing image and the predicted semantics, which preserves local warping similarity. Finally, a canonical try-on synthesizer fuses all the processed information to generate the clothed person image. Extensive experiments on conventional benchmarks along with user studies demonstrate that our framework achieves state-of-the-art performance both quantitatively and qualitatively. Notably, RT-VTON shows compelling results on a wide range of non-standard clothes.* *Project page:* https://lzqhardworker.github.io/RT-VTON/.

## 1. Introduction

Virtual try-on is a rapidly advancing topic in both academia and industry with the increasing power of gener-

✉ Corresponding author

ative models. Various pipelines [4, 10, 14, 37, 42] are proposed to build the system, but it remains challenging to perform full-range try-on with different clothing types in real-world scenarios. Standard clothes such as T-shirts and long-sleeve jackets show clear relationship with the reference person, while non-standard clothes can involve irregular patterns and design, thus resulting in more ambiguous corresponding relations. Two typical non-standard types are the off-shoulder clothes (normal collar with shoulder exposed) and the word-shoulder clothes (a horizontal collar line towards shoulder). The results on those kinds of non-standard clothes are scarcely reported in any of the try-on papers [4, 10, 14, 29, 37, 42].

Earlier works [14, 37] utilize the coarse shape and pose map to synthesize try-on results with Thin-plate Spline (TPS) warping. Pioneering methods [10, 17, 42] ameliorate the blurry artifacts caused by coarse shape [14, 37] by firstly predicting the semantic layout with the target clothing image and then warping the clothing image by regularized TPS, producing better results with sharper boundaries. However, these methods [10, 42] still struggle to precisely depict the "after-try-on" semantics, where the functioning zone is restricted to standard clothes. Another barrier preventing full-range virtual try-on is the misalignment of the clothing image with the reference person. TPS is a usual practice as used in [2, 4, 14, 37, 42] to spatially transform the clothing image while preserving the characteristics. However, over-distortion of the clothing image hinders the TPS-based methods, instigating the increasing preference for affine-based algorithms [9,20]. As opposed to TPS, affine-based methods [9, 20] demonstrate large potential in generating undistorted results, but the non-rigid part of deformation is not involved which fails to mimic the natural interaction between the clothes and the person. Flow-based methods [5, 11, 12, 41] embed the maximized capacity in the deformation modeling which densely predict the pixel-wise offset field. However, without the ground-truth flow, optimizing the flow network is only possible with strong regularization priors such as affine prior, total variance constraint, or second-order Laplacian penalty.

To achieve full-range virtual try-on, we propose a principled framework, Recurrent Tri-Level Transform (RT-VTON) which deeply mines the "after-try-on" semantics by accurately predicting the semantic layout of the reference person given the target clothing image, and then coherently deform the clothing image with our semi-rigid deformation to balance rigidity and flexibility. Specifically, RT-VTON follows a conventional split-transform-merge scheme (Fig. 2) as in [4, 5, 22, 41, 42]. The first module is Semantic Generation Module (SGM), which gradually transforms the tri-level feature codes to predict the semantic segmentation of the body parts and the clothing region. As opposed to prior works, our SGM can accurately capture the correlation between the target clothing image with the human body, and thus perform full-range try-on especially for non-standard clothes (see Fig. 1). The second module is Clothes Deformation Module (CDM) which applies a novel semi-rigid deformation to align the target clothing image according to the semantic output of SGM. We borrow the widely-used geometric editing technique from graphics [16, 34] and, for the first time, integrate it within a differentiable learning-based framework. Finally, a Try-on Synthesizer Module (TOM) similar to [4, 42] fuses the semantic segmentation, the warped clothes as well as the non-target body image to synthesize the final try-on output, where an auxiliary clothes reconstruction loss is used to enhance texture preserving.

We summarize our contributions as follows. **1)** We propose a new image-based virtual try-on framework, *i.e*. RT-VTON, which accurately depicts the "after-try-on" semantics and thus greatly improve try-on quality and adaptability for full-range garment types. **2)** A novel Recurrent Tri-Level Transform is proposed to improve the semantic layout prediction, which gradually updates three different levels of clothing representations, namely clothes code, pose code, and parsing code, by local gated attention mechanism with non-local correspondence learning. **3)** To perform undistorted clothes warping, we design a semi-rigid deformation to align the clothing image with the predicted semantics, which preserves local warping similarity. **4)** Extensive experiments demonstrate that the proposed method can perform realistic virtual try-on for both standard and non-standard clothes, outperforming the state-of-the-art methods qualitatively and quantitatively.

## 2. Related Works

**Fashion Analysis and Synthesis.** Recently, fashion-related topics become increasingly popular due to their potential capability to change our life. Clothing attribute recognition and prediction [24, 36] attracts much attention to automatically understand the clothing semantics. Landmark detection [15, 19, 25, 39] is another fast-growing area that is fundamental for other fashion-related applications. With the help of powerful Generative Adversarial Networks (GANs), Fashion image synthesis [1, 13, 23] is another popular field attracting both researchers and companies.

**Pose-Guided Person Image Generation.** Pose-guided image generation aims at synthesizing photo-realistic person images with specified poses, given target poses and reference person images, which is first introduced in $PG^2$ [28]. $PG^2$ utilizes a two-stage image-to-image translation network to solve this task. Later, variational U-net [8] combines U-net and Conditional VAE [35] to disentangle appearance and pose. However, spatial information is ignored in these methods, which leads to appearance misalignment. PATN [45] utilizes pose information in progressive atten-

tion modules, which possesses the appearance coherence and shape consistency with the input images. GFLA [31] learns pixel-wise flow and utilizes local attention to warp the source person image. Recently, SPGNet [27] , a new two-stage method, utilizes pose and semantic information to guide person image generation.

**Image-based Virtual Try-on.** Virtual try-on aims at generating photo-realistic person in specified clothing image, given target clothes and the reference person image. Recently, methods based on deep learning, especially Generative Adversarial Networks (GANs), have received considerable attention. Generally, virtual try-on methods based on deep learning can be categorized as 3D-based method and 2D-based method. Since it is difficult to collect 3D try-on data, 2D methods are more widely discussed in academia. VITON [14], CP-VTON [37] utilize coarse shape of the body, pose map, and TPS-based warping method to deform the clothes and generate a person wearing specified clothing image. ACGPN [42] proposes a split-transform-merge scheme to generate the try-on image by adaptively generating and preserving image contents, achieves photo-realistic results. DCTON [10] proposes a method following cycle consistency learning, which stabilizes the try-on image synthesis but may mis-preserve the image contents from the reference person. PF-AFN [11] is a two-staged model that adopts knowledge distillation to correct the error in semantic parsing. The distillation trick is definitely a viable post-processing for our method, but the try-on quality is still largely dependent on the first one-pass try-on stage, which is our main focus.

## 3. Recurrent Tri-Level Virtual Try-On

**Framework Overview.** To generate photo-realistic try-on results, Recurrent Tri-Level Transform (RT-VTON) follows a split-transform-merge scheme, consisting of three modules: Semantic Generation Module (SGM), Clothes Deformation Module (CDM), and Try-on Synthesizer Module (TOM). We firstly remove the face, upper clothes, and arm labels from the input parsing to derive partial parse; the original clothing shape is thus agnostic to the network. Then SGM predicts the "after-try-on" semantic layout given the target clothes as well as the reference pose map. With accurate semantic segmentation, we can adaptively generate and preserve the image contents by computing the intersection of skin regions, *i.e*. residual body. Our key insights towards the framework are SGM for semantics transfer and the CDM for geometry transfer.

### 3.1. Recurrent Tri-Level Transform

Prior works such as CP-VTON [37] and VITON [14] use coarse body shape as input instead of semantic segmentation, losing the capability to grasp the fine details of clothing and non-clothing areas. ACGPN [42] is the first to build the semantic-based pipeline to generate photo-realistic results but fails to stabilize the semantic prediction process. Besides, due to the misalignment of the clothing image and the reference person, it remains a big challenge to accurately retain the clothing shape after the try-on process. To address this problem, our Recurrent Tri-Level Transform is based on three levels of clothing representations, namely clothes code, pose code, and parsing code. Motivated by the real clothes-wearing process of human action, we try to mimic this process which firstly finds out the long-range correspondence and then generates the semantic layout. This breaks the traditional pipeline that directly learns the semantic transformation conditioned on the clothing image as well as the pose map, failing to predict the accurate semantic layout, especially for non-standard clothes as in Fig. 4. Our pipeline combines the local gated attention with global correspondence learning to gradually refine the tri-level feature codes, which empowers our SGM to predict the accurate semantic layout for further generation.

**Local Gated Attention**. We model the local gated attention mechanism as a self-correcting process that filters the irrelevant features. The dual attention masks are computed respectively for the pose code and clothes code from the parsing code.

Starting form the initial parsing code $F_0^S$, clothes code $F_0^C$ and pose code $F_0^P$, the $t^{th}$ pose code $F_t^P$ and the $t^{th}$ clothes code $F_t^C$ are updated by the attention masks from $t^{th}$ parsing code $F_t^S$. The design of the gated blocks follows the conventional structure as [28,43]. The widely used entry-wise $sigmoid$ gating is applied here, formulated as:

$$\begin{aligned} M_t^C &= \sigma(\text{conv}_{S\to C}(F_{t-1}^S)), \\ M_t^P &= \sigma(\text{conv}_{S\to P}(F_{t-1}^S)), \end{aligned} \tag{1}$$

where the $\text{conv}_{S\to C}$, $\text{conv}_{S\to P}$ indicate the convolutional layers. The $\sigma$ indicates the entry-wise $sigmoid$ function.

**Non-Local Correspondence**. We try to find the correlation between the clothes code $F_t^C$ and the pose code $F_t^P$ in $t^{th}$ block with the correspondence layer proposed in [44]. Specially, $F_t^P$ and $F_t^C$ are downsampled with convolutional layers to extract the high-level features, which are later flattened into $x_t'^P \in \mathbb{R}^{HW\times C}$ (pose), and $x_t'^C \in \mathbb{R}^{HW\times C}$ (clothes), and the correlation matrix $\mathcal{M}_C^t \in \mathbb{R}^{HW\times HW}$ is computed by pair-wise feature correlation,

$$\mathcal{M}_C^t(u,v) = \frac{\hat{x}_t^C(u)^T \hat{x}_t^P(v)}{\left\|\hat{x}_t^C(u)\right\| \left\|\hat{x}_t^P(v)\right\|}, \tag{2}$$

where $\hat{x}_t^C(u)$ and $\hat{x}_t^P(v)$ indicate the channel-wise centralized feature of $x_t'^C$ and $x_t'^P$. $\mathcal{M}_C^t(u,v)$ represents the corresponding similarity, in the $t^{th}$ block.

The non-local correspondence matrix $\mathcal{M}_C^t$ is then used to transform the flattened clothes code $x_t^C$ from $F_t^C$ by

$$\bar{x}_t^C = softmax_v(\alpha \mathcal{M}_C^t) x_t^C, \tag{3}$$

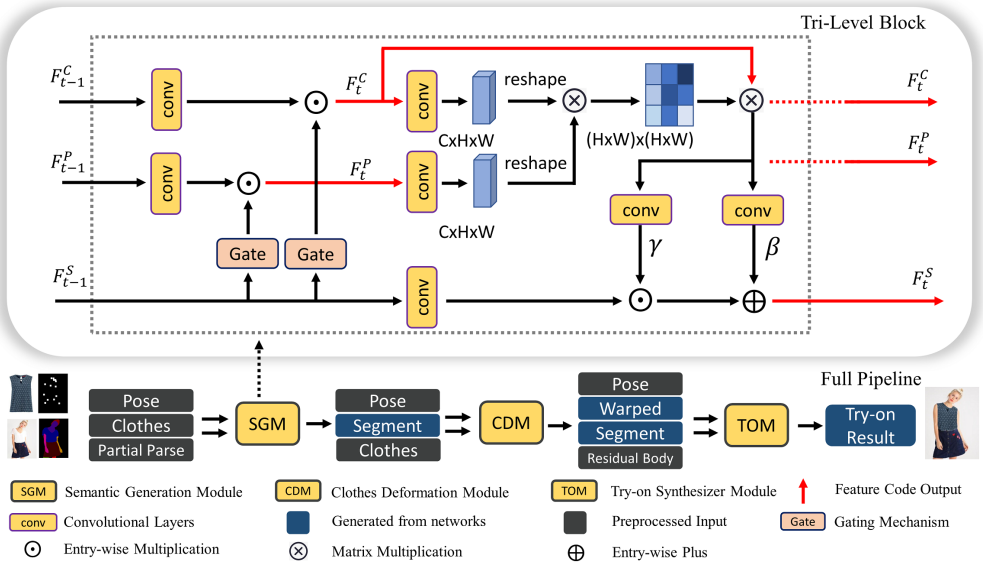



Figure 2. The overall pipeline of our Recurrent Tri-Level Transform (RT-VTON) as well as the detailed tri-level feature codes updating scheme. The framework firstly removes the face, upper clothes and arm labels from the input parsing to get partial parse, and then we predict the "after-try-on" semantic layout given the target clothes. By computing the intersection of skin regions, the unchanged body texture, *i.e.* residual body, can be extracted as the input of TOM.

where $\alpha$ is a sharpening parameter as used in [44], $x_t^C$ is unfolded with sliding window in actual implementation, and $softmax_v$ is the $softmax$ operation along the row dimension. Then we reshape the flattened transformed clothes code $\bar{x}_t^C$ back to get $\bar{F}_t^C$.

**Code Update**. With the computed attention masks $M_t^C$ and $M_t^P$, the clothes code $F_t^C$ and the pose code $F_t^P$ are updated by:

$$\begin{aligned} F_t^C &= M_t^C \odot \text{conv}_C(F_{t-1}^C) + F_{t-1}^C \\ F_t^P &= M_t^P \odot \text{conv}_P(F_{t-1}^P) + F_{t-1}^P, \end{aligned} \tag{4}$$

where $\odot$ denotes entry-wise multiplication. Then the high-level features are extracted to compute the correlation matrix $\mathcal{M}_C^t$, and by applying Eq. 3 we have the transformed clothes code $\bar{x}_t^C$. We update the parsing code $F_t^S$ by:

$$F_t^S = \gamma(\bar{F}_t^C) \odot F_{t-1}^S + \beta(\bar{F}_t^C), \tag{5}$$

where $\gamma(\cdot)$ and $\beta(\cdot)$ indicate the conditional scale and offset parameter computation, following the Spatial Feature Transform (SFT) [40] design. Instead of directly relying on the clothing feature, the parsing code is modulated by the spatially transformed clothes code $\bar{F}_t^C$, that effectively bridges the misalignment between the target clothes and the reference person. Reconstruction loss is added to help the correspondence learning by deforming the downsampled clothing image, with the clothes on the reference person as the ground-truth. Fig. 6 demonstrates the effectiveness of non-local correspondence learning to help understand the "after-try-on" semantics.

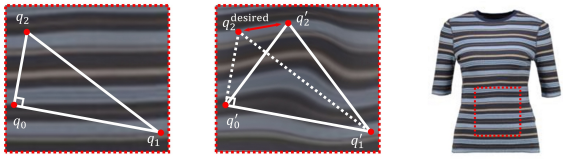
### 3.2. Semi-Rigid Deformation

Having predicted the semantic layout of the "after-try-on" person, we can deform the clothing image to transfer the texture. The common practice [10, 14] uses Thin-plate Spline (TPS) [7] to model the spatial deformation. Motivated by the collinearity of affine transformation, ACGPN [42] proposes a second-order difference constraint to penalize the non-affine part of TPS warping. DC-TON [10] proposes a homography regularization to stabilize the TPS training, but over-distortion can still be observed as in Fig. 5. Previous methods try to combine the flexibility of TPS with the rigidity of affine transformation but fail to find an equilibrium of this trade-off.

To address this problem, we propose a semi-rigid deformation that models the deformation as a learnable Moving Least Squares [34] problem to balance the trade-off of flexibility and rigidity. The influence of the control points decays quadratically along with the distance, therefore allowing local flexibility while computing the individual affine transformation parameters for each point. We give a clear explanation of the Moving Least Squares problem in the supp. materials. We define the uniformly sampled initial control points as $q$ and the predicted target points as $q'$. Given a point $v$ in the image, we compute different affine transformations for each $v$ by applying the decaying weights

$$w_i = \frac{1}{|q_i' - v|^{2\alpha}}, \tag{6}$$

where $\alpha$ is a decay parameter with the default value 1 and $i$ denotes the $i^{th}$ point. By solving the Least Squares prob-




Initial Triangle Deformed Triangle Target Clothes

Figure 3. Illustration of computing As-Similarity-As-Possible (ASAP) [16] regularization. The warping effect here is only for easier understanding, which is not the actual deformation result.

lem, we can perform the semi-rigid deformation on our target clothing image according to the predicted control points $q'$. Via applying the spatially decaying weights, the advantage of affine transformation is fully utilized while allowing local flexibility.

**Local Similarity Preservation**. We propose imposing an As-Similarity-As-Possible (ASAP) [16, 26] constraint which is computed within each quad of the predicted control points $q'$. Every quad is split into two triangles. As shown in Fig. 3, ASAP constraint enforces similarity transform of each triangle by penalizing the deformed triangle $\{q_0', q_1', q_2'\}$. We define the relative coordinates $q_1\,\{x_{01}, 0\}$ and $q_2\,\{0, y_{02}\}$, where $x_{01}$ and $y_{02}$ are the fixed intervals of uniformly sampled control points. Then $q_2$ can be defined by $q_0$ and $q_1$as:

$$q_2 = q_0 + \frac{y_{02}}{x_{01}} R_{90} \overrightarrow{q_0 q_1}, \tag{7}$$

where $R_{90}$ indicates counterclockwise rotation of 90 degrees. Given $q_0'$ and $q_1'$ we can compute the desired position of $q_2'$:

$$q_2^{\text{desired}} = q_0' + \frac{y_{02}}{x_{01}} R_{90} \overrightarrow{q_0' q_1'}, \tag{8}$$

and our regularization term can be formulated as:

$$E^{\{q_2'\}} = \left\| q_2^{\text{desired}} - q_2' \right\|^2, \tag{9}$$

where $E^{\{q_2'\}}$ denotes the error term for $q_2'$. Similarly we can apply the same mechanism to $q_1'$ for the other triangle in the same quad,

$$E^{\{q_1', q_2'\}} = \sum_{i \in \{1,2\}} \left\| q_i^{\text{desired}} - q_i' \right\|^2, \tag{10}$$

where the error for each quad is summed up to form the final regularization loss. L1 loss and perceptual loss [18] are applied to guide the image warping, regularized by the ASAP constraint. To this end, we have successfully prepared all the ingredients for the final synthesis.

### 3.3. Try-on Synthesizer

Given the predicted semantic layout as well as the deformed clothing image, the proposed Try-on Synthesizer Module (TOM) generates the clothed person with the above input. We adopt a similar adaptive generation and preservation strategy as addressed in [42] to preserve the non-target body parts while generating the exposed body texture. To encourage the network to preserve the warped clothing texture, RT-VTON reconstructs the deformed clothing input at the same time, which helps the network encode more identity mapping clues to preserve the original characteristics.

To train the adaptive preservation and generation, mask inpainting strategy [42] is applied by randomly removing the body parts to build the capability of generating the missing skin while preserving the unchanged pixels (residual body). We use the masks offered from Irregular Mask Dataset [21] to randomly remove the face, neck, and arms while training. During the training, our TOM generates a triplet $\{I_S', C_R, \alpha\}$, where $I_S'$ is the generated clothed body, $C_R$ is the reconstructed warped clothes for auxiliary supervision, and $\alpha$ is a composition mask to composite the generated image with the warped clothes $C_W$ by

$$I_S = \alpha \odot I_S' + (1 - \alpha) \odot C_W, \tag{11}$$

where $\odot$ denotes entry-wise multiplication and $I_S$ is our syntheiszed try-on result.

L1 loss, perceptual loss [18] and adversarial loss are applied for the generation of the clothed body as well as the warped clothes.The same regularization is applied to $\alpha$ following [37].

During testing, with the semantic layout of the reference person and the predicted semantic layout, we can fully preserve the unchanged skin pixels by feeding the residual body $I_R$ to the TOM, defined as

$$I_R = I \odot M_{skin} \odot M'_{skin}, \tag{12}$$

where $\odot$ denotes the entry-wise multiplication, $I$ is the input reference person, $M_{skin}$ is the skin region of the reference person, and $M'_{skin}$ is the skin area of the predicted semantic layout.

## 4. Experiments

### 4.1. Datasets and Comparisons

Experiments are conducted on the standard virtual try-on benchmark (*i.e.*, VITON dataset) containing about 19,000 image pairs, each of which includes a reference person image and a corresponding target clothing image. After removing the invalid image pairs following [37], it yields 16,253 pairs, resulting in a training set of 14,221 pairs and a testing set of 2,032 pairs.

**Non-standard Clothes Set.** For quantitative comparison, we try to exhaust the non-standard clothes in the test set, resulting in a non-standard clothes set with 48 clothing images by manual selection, including the off-shoulder clothes, the word-shoulder clothes, as well as the clothes with complex


Figure 4. Visual comparison of four virtual try-on methods in a standard to non-standard manner (top to bottom). Four methods perform quite well for the standard cases without shape changes, but fail drastically when performing large shape transformation towards non-standard cases. With our Tri-Level Transform and semi-rigid deformation, RT-VTON produces photo-realistic results for the full-range of clothing types and preserves the fine details of the clothing texture.

laces. Typical non-standard and standard clothes are presented in Fig. 1 and the differences between standard and non-standard clothes are obvious to understand by comparison. Due to the scarcity of non-standard clothes, the non-standard statistics are only for reference.

**Comparisons.** RT-VTON is compared with three state-of-the-art methods including CPVTON+ [29], ACGPN [42], and DCTON [10] with official implementations.

### 4.2. Experimental Setups

**Network Architectures.** RT-VTON consists of three modules, SGM, CDM, and TOM. SGM is composed of Tri-Level Blocks. The design of feature extraction and downsampling before computing the non-local correspondence matrix follows [44] for robust correspondence modeling. Notably, the extracted high-level feature can also be used as the updated clothes code after upsampling for the next block, which is taken as our actual implementation. CDM is similarly designed as [37], which is a conventional structure. We use Res-Unet as [11] to build up our TOM to preserve the input information. All the images are in resolution $256 \times 192$. Gaussian pose heatmap and Gaussian pose segment map [27] are used in SGM and TOM.

**Training Details.** Random flipping augmentation of the target clothes is applied in training SGM. The SGM and CDM are trained respectively, and the TOM is trained with the warped clothing images from a pretrained CDM. We use the same ResNet-based discriminator as [31]. SGM and TOM are trained with batchsize 4 for 20 epochs while CDM is trained with batchsize 1 for 20 epochs. ASAP regularization weight in the semi-rigid deformation training is set as 0.001. Learning rate is initialized as 0.0002 and we adopt Adam optimizer with the default hyper-parameters. All the codes are implemented in PyTorch and trained on 1 Tesla V100 GPU.

**Testing and Evaluation Metrics.** Hand segmentation is pasted back in the test phase to better preserve the finger details. To prepare the test pairs, for each reference person we randomly assign a target clothing within the dataset partition. For the full dataset testing when computing the FID score, we randomly shuffle the human-clothes pairs to make sure each clothing image is assigned once. For the non-standard (N.S.) setting, each reference person is randomly assigned with a non-standard clothing image.

### 4.3. Qualitative Results

**Try-On Comparisons.** We conduct a visual comparison experiment in Fig. 4 with three methods, including the state-of-the-art semantic-based method DCTON [10], together with ACGPN [42] and CPVTON+ [29]. We can see that CPVTON+ can only produce blurry results with poor body part textures, while ACGPN generates a clear and sharp clothed body. However, ACGPN still fails to generate accurate semantic layout according to the given target clothing image, especially for non-standard clothes. DCTON improves the semantic consistency, as in Fig. 4 (a, i), which successfully produces a person wearing long-sleeve clothes while ACGPN wrongly preserves the structure of the original short-sleeve reference person. However, mis-preserved sleeve borders can be seen in results from DCTON in Fig. 4 (a c, i), which largely harms the visual quality. On the contrary, RT-VTON can generate accurate "after-try-on" se-


Figure 5. The visual comparison of the image deformation methods between the TPS warping and our semi-rigid deformation. State-of-the-art TPS-based try-on method DCTON is chosen as our baseline to evaluate the effectiveness of our method.

mantics regardless of the original clothes on the reference person (as in Fig. 4 (a, i)). Since the clothing part of the reference person is removed in our method, RT-VTON is invulnerable to the mis-perserving problem like DCTON.

When compared within the non-standard clothes setting, RT-VTON can give even better try-on outputs. In Fig. 4 (b, i), the laces of this sleeveless shirt are accurately preserved, while ACGPN recognizes it as a short-sleeve and DCTON also fails to preserve the fine details. From the results in non-standard clothes, we can see that the existing methods are unable to generate photo-realistic results while retaining the accurate clothing shape. In Fig. 4 (c-d, i) and (b-d, ii), only CPVTON+ can somehow preserve the off-shoulder characteristics, while ACGPN and DCTON generate either standard short-sleeve or long-sleeve results, overfitting the standard clothing shapes. Without using semantic segmentation, CPVTON+ directly borrows the structure from the warped clothes which helps retain the off-shoulder characteristics but also fails to generate clear clothes-body boundaries. We can see the trade-off of using semantic layout comparing CPVTON+ with ACGPN and DCTON, that semantic layout helps greatly in improving image quality as well as preserving the non-target body part details, but at the same time, an incorrect segmentation may lead to unpredictable artifacts. RT-VTON breaks this trade-off by using the Tri-Level Transform which produces even better structures than non-semantic methods (CPVTON+) without losing the advantages of the semantic-based pipelines (ACGPN, DCTON).

**Geometry Deformation Comparisons**. We also conduct qualitative experiments on the effectiveness of semi-rigid deformation. To demonstrate the superiority over the conventional TPS-based methods, we pick the best TPS-based method, *i.e*. DCTON as our baseline. As in Fig. 5, we can see that our semi-rigid deformation can accurately align the clothing image while preserving the local similarity. The clothes deformed by DCTON may be over-distorted near the boundary, as in the fourth column. And we can see the non-uniform squeezing of the off-shoulder clothes in the fifth column, which may be caused by the difficulty of understanding the clothes-human relations. From the second column, we can also see the inherent flexibility of our semi-rigid deformation by computing the Moving Least Squares affine parameters.

Table 1. Quantitative Comparisons. "N.S." denotes non-standard. We show the Fréchet Inception Distance (FID) [30] and user study results of four methods. FID is the lower the better. User study is given by the preference ratio for our method, which is the higher the better. '-' indicates the placeholder. Without the official implementations of [5,12,17], we give an extra comparison for reference with their reported numbers in the supp. materials.

| Method | FID | | User Preference | |
|---|---|---|---|---|
| | overall | N.S. | overall | N.S. |
| CPVTON+ [37] | 21.29 | 24.10 | 82.35% | 82.17% |
| ACGPN [42] | 16.46 | 19.22 | 73.88% | 75.00% |
| DCTON [10] | 16.37 | 20.42 | 68.87% | 71.67% |
| RT-VTON | **11.66** | **17.24** | - | - |

## 4.4. Quantitative Results

Quantitative evaluation of try-on task is hard to conduct as there is no ground-truth of the reference person in the target clothes. The Fréchet Inception Distance (FID) [30] is adopted to measure the similarity of data distribution between the generated results and the reference data. Since Inception Score (IS) [33] is only effective in dataset similar to ImageNet [6] as addressed in [3], we do not adopt IS as our metric to evaluate virtual try-on.

The quantitative results are given in Tab. 1. RT-VTON achieves the state-of-the-art results in both overall and non-standard settings by a large margin. In particular, RT-VTON outperforms CPVTON+, ACGPN and DCTON by 9.63, 4.80 and 4.71 respectively in the overall setting. We can see the large gap between non-semantic method, CPVTON+ with ACGPN and DCTON. The FID scores of ACGPN and DCTON are indistinguishable, partly indicating the same structure limitations that both methods suffer from.

## 4.5. User Study

Image metric may have limitations in depicting the try-on quality. To further demonstrate the superiority of our method, we conduct a user study for the whole test set and also apply for a non-standard setting. 25 volunteers are invited to our user study. 30 image pairs from either settings (overall, non-standard) are assigned for each volunteer, containing a reference person, a target clothing image, a result from RT-VTON and a result from randomly selected baseline methods. To improve test accuracy, the results of two methods are random shuffled so the users cannot tell from the position to prevent casual gaming. From user study in Tab. 1, RT-VTON outperforms the existing state-of-the-art methods compellingly in both overall and non-standard set-


Figure 6. Visual ablation study of Semantic Generation Module (SGM) in RT-VTON. Tri-Level Transform is compared with plain encoder-decoder [38] and Unet [32] structure.

Table 2. Ablation study on Semantic Generation Module (SGM). We compare them in a segmentation reconstruction setting, and four categories "face, left arm, right arm, upper clothes" are taken into consideration. Mean IoU (Intersection over Union) is adopted, which is the higher the better.

| Different configuration of SGM | Mean IoU (%) |
|---|---|
| A plain encoder-decoder architecture [38] | 86.31 |
| A conventional Unet backbone used in [4, 42] | 86.77 |
| Tri-Level Transform (ours) | **88.11** |

tings. Our proposed framework helps both the overall and the non-standard setting, which explains the close preference ratios in two settings.

### 4.6. Ablation Study

Our ablation studies are conducted mainly on analyzing the effectiveness of our Tri-Level Block in Semantic Generation Module (SGM). Three settings are given as: **1)** full RT-VTON with Tri-Level Transform, **2)** RT-VTON with plain encoder-decoder connected by residual blocks, following [38], **3)** RT-VTON with Unet [32] as SGM, which is a common backbone in designing the try-on pipelines [4, 42]. Mean Intersection over Union (IoU) metric is used to evaluate the semantic prediction for the same clothes-person pairs, as in Tab. 2. Unpaired try-on results are also visualized in Fig. 6; we can see clearly that Unet or encoder-decoder based SGMs are unable to capture the complex shapes of non-standard clothes; they recognize the word-shoulder shirt as a usual U-collar shirt. Moreover, the sling vest (second row) is mistakenly modeled as long-sleeve shirt due to the uninformative shape clues of the target clothes. Our full model can combine the advantages of long-range modeling and local attention to fully utilize the clothes-human correlation and thus successfully depict the accurate semantic layout even for non-standard clothes.

**Effectiveness of Non-Local Correspondence.** Since the correlation matrix is computed in a downsampled feature space ($16 \times 12$), we present a patch-based correspondence in image level by manually selected locations in Fig. 7. It is clear that the non-local correspondence learning helps capture the non-standard clothing pattern (on the left), which demonstrates strong relationship of the off-shoulder area to retain the clothing shape. Moreover, the boundaries of the sleeves (on the right) are well depicted with the target clothes which leverages the long-range correlation to reconstruct the final semantic layout.

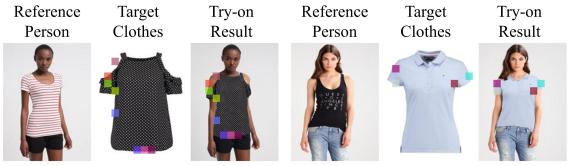



Figure 7. Visualization of our non-local correspondence given some manually selected positions. Since the correlation matrix is computed in downsampled feature space, the corresponding points are shown in patches with the same color.

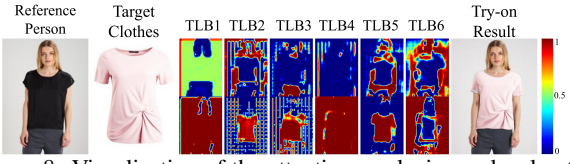



Figure 8. Visualization of the attention masks in our local gating mechanism for clothes code (top) and pose code (bottom). TLB1-6 denotes the six Tri-Level Blocks we use in our Semantic Generation Module (SGM).

**Effectiveness of Gated Attention.** In Fig. 8, we extract the attention masks for the six Tri-Level Blocks used in RT-VTON. The first row shows the masks for the clothes code and the bottom row covers the masks for the pose code. The gated attention transfers the feature in a gradual manner, which coincides nicely with the human intuition. The initial masks are mixtures of input segmentation, but the clothing shapes and sleeve boundaries are gradually revealed as the parsing code is modulated according to the target clothes. In the last two columns, we can see the clear sleeve boundary of the "after-try-on" person which helps determine the target clothing shape.

## 5. Conclusion

In this work, we propose a novel Recurrent Tri-Level Transform (RT-VTON), which embeds two principled insights in semantics transfer as well as geometry transfer: **1)** Tri-Level Transform which models the long-range dependency with local gated attention to predict accurate semantic layout; **2)** semi-rigid deformation which tries to balance the trade-off of rigidity and flexibility in clothes warping.

**Acknowledgements.** This work is supported by NTU NAP, MOE AcRF Tier 1 (2021-T1-001-088), and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

## References


[1] Kenan Emir Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A. Kassim. Semantically consistent hierarchical text to fashion image synthesis with an enhanced-attentional generative adversarial network. In *ICCV Workshops*, pages 3121–3124. IEEE, 2019. 2

[2] Kumar Ayush, Surgan Jandial, Ayush Chopra, Mayur Hemani, and Balaji Krishnamurthy. Robust cloth warping via multi-scale patch adversarial loss for virtual try-on framework. In *ICCV Workshops*, pages 1279–1281. IEEE, 2019.

[3] Shane T. Barratt and Rishi Sharma. A note on the inception score. *CoRR*, abs/1801.01973, 2018. 7

[4] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. VITON-HD: high-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, pages 14131–14140. Computer Vision Foundation / IEEE, 2021. 2, 8

[5] Ayush Chopra, Rishabh Jain, Mayur Hemani, and Balaji Krishnamurthy. Zflow: Gated appearance flow-based virtual try-on with 3d priors. *CoRR*, abs/2109.07001, 2021. 2, 7

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009. 7

[7] Jean Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive Theory of Functions of Several Variables*, volume 571 of *Lecture Notes in Mathematics*, pages 85–100. Springer, 1976. 4

[8] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, pages 8857–8866. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[9] Matteo Fincato, Federico Landi, Marcella Cornia, Fabio Cesari, and Rita Cucchiara. VITON-GT: an image-based virtual try-on model with geometric transformations. In *ICPR*, pages 7669–7676. IEEE, 2020. 2

[10] Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, and Ping Luo. Disentangled cycle consistency for highly-realistic virtual try-on. In *CVPR*, pages 16928–16937. Computer Vision Foundation / IEEE, 2021. 2, 3, 4, 6, 7

[11] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *CVPR*, pages 8485–8493. Computer Vision Foundation / IEEE, 2021. 2, 3, 6

[12] Xintong Han, Weilin Huang, Xiaojun Hu, and Matthew R. Scott. Clothflow: A flow-based model for clothed person generation. In *ICCV*, pages 10470–10479. IEEE, 2019. 2, 7

[13] Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R. Scott, and Larry Davis. Finet: Compatible and diverse fashion image inpainting. In *ICCV*, pages 4480–4490. IEEE, 2019. 2

[14] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. VITON: an image-based virtual try-on network. *CoRR*, abs/1711.08447, 2017. 2, 3, 4

[15] Chang-Qin Huang, Jikai Chen, Yan Pan, Hanjiang Lai, Jian Yin, and Qionghao Huang. Clothing landmark detection using deep networks with prior of key point associations. *IEEE Trans. Cybern.*, 49(10):3744–3754, 2019. 2

[16] Takeo Igarashi, Tomer Moscovich, and John F. Hughes. As-rigid-as-possible shape manipulation. *ACM Trans. Graph.*, 24(3):1134–1141, 2005. 2, 5

[17] Surgan Jandial, Ayush Chopra, Kumar Ayush, Mayur Hemani, Abhijeet Kumar, and Balaji Krishnamurthy. Sievenet: A unified framework for robust image-based virtual try-on. In *WACV*, pages 2171–2179. IEEE, 2020. 2, 7

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV (2)*, volume 9906 of *Lecture Notes in Computer Science*, pages 694–711. Springer, 2016. 5

[19] Hyo Jin Kim, Doo Hee Lee, Asim Niaz, Chan Yong Kim, Asif Aziz Memon, and Kwang Nam Choi. Multiple-clothing detection and fashion landmark estimation using a single-stage detector. *IEEE Access*, 9:11694–11704, 2021. 2

[20] Kedan Li, Min Jin Chong, Jingen Liu, and David A. Forsyth. Toward accurate and realistic virtual try-on through shape matching and multiple warps. *CoRR*, abs/2003.10817, 2020. 2

[21] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV (11)*, volume 11215 of *Lecture Notes in Computer Science*, pages 89–105. Springer, 2018. 5

[22] Guoqiang Liu, Dan Song, Ruofeng Tong, and Min Tang. Toward realistic virtual try-on through landmark guided shape matching. In *AAAI*, pages 2118–2126. AAAI Press, 2021. 2

[23] Jingyuan Liu and Hong Lu. Deep fashion analysis with feature map upsampling and landmark-driven attention. In *ECCV Workshops (3)*, volume 11131 of *Lecture Notes in Computer Science*, pages 30–36. Springer, 2018. 2

[24] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, pages 1096–1104. IEEE Computer Society, 2016. 2

[25] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild. In *ECCV*, 2016. 2

[26] Ziwei Liu, Lu Yuan, Xiaoou Tang, Matt Uyttendaele, and Jian Sun. Fast burst images denoising. *TOG*, 2014. 5

[27] Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. In *CVPR*, pages 10806–10815. Computer Vision Foundation / IEEE, 2021. 3, 6

[28] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NIPS*, pages 406–416, 2017. 2, 3

[29] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul Rosin, and Yu-Kun Lai. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2, 6

[30] Julianna Pinele, João E. Strapasson, and Sueli I. R. Costa. The fisher-rao distance between multivariate normal distributions: Special cases, boundsand applications. *Entropy*, 22(4):404, 2020. 7

[31] Yurui Ren, Ge Li, Shan Liu, and Thomas H. Li. Deep spatial transformation for pose-guided person image generation and animation. *CoRR*, abs/2008.12606, 2020. 3, 6

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI (3)*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015. 8

[33] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *CoRR*, abs/1606.03498, 2016. 7

[34] Scott Schaefer, Travis McPhail, and Joe D. Warren. Image deformation using moving least squares. *ACM Trans. Graph.*, 25(3):533–540, 2006. 2, 4

[35] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *NIPS*, pages 3483–3491, 2015. 2

[36] Li Sun, Gerardo Aragon-Camarasa, Simon Rogers, Rustam Stolkin, and J. Paul Siebert. Single-shot clothing category recognition in free-configurations with application to autonomous clothes sorting. *CoRR*, abs/1707.07157, 2017. 2

[37] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV (13)*, volume 11217 of *Lecture Notes in Computer Science*, pages 607–623. Springer, 2018. 2, 3, 5, 6, 7

[38] 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*, 2018. 8

[39] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *CVPR*, pages 4271–4280. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[40] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 4

[41] Zhenyu Xie, Xujie Zhang, Fuwei Zhao, Haoye Dong, Michael C. Kampffmeyer, Haonan Yan, and Xiaodan Liang. WAS-VTON: warping architecture search for virtual try-on network. In *ACM Multimedia*, pages 3350–3359. ACM, 2021. 2

[42] 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 *CVPR*, pages 7847–7856. Computer Vision Foundation / IEEE, 2020. 2, 3, 4, 5, 6, 7, 8

[43] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-form image inpainting with gated convolution. In *ICCV*, pages 4470–4479. IEEE, 2019. 3

[44] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *CVPR*, pages 5142–5152. Computer Vision Foundation / IEEE, 2020. 3, 4, 6

[45] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *CVPR*, pages 2347–2356. Computer Vision Foundation / IEEE, 2019. 2