ClothFormer: Taming Video Virtual Try-on in All Module

Jianbin Jiang[1*], Tan Wang[2], He Yan[2], Junhui Liu[2]
[1]BIGO [2]iQIYI Inc.
jiangjianbin@bigo.sg, {wangtan, yanhe, liujunhui}@qiyi.com

Abstract

The task of video virtual try-on aims to fit the target clothes to a person in the video with spatio-temporal consistency. Despite tremendous progress of image virtual try-on, they lead to inconsistency between frames when applied to videos. Limited work also explored the task of video-based virtual try-on but failed to produce visually pleasing and temporally coherent results. Moreover, there are two other key challenges: 1) how to generate accurate warping when occlusions appear in the clothing region; 2) how to generate clothes and non-target body parts (e.g. arms, neck) in harmony with the complicated background; To address them, we propose a novel video virtual try-on framework, ClothFormer, which successfully synthesizes realistic, harmonious, and spatio-temporal consistent results in complicated environment. In particular, ClothFormer involves three major modules. First, a two-stage anti-occlusion warping module that predicts an accurate dense flow mapping between the body regions and the clothing regions. Second, an appearance-flow tracking module utilizes ridge regression and optical flow correction to smooth the dense flow sequence and generate a temporally smooth warped clothing sequence. Third, a dual-stream transformer extracts and fuses clothing textures, person features, and environment information to generate realistic try-on videos. Through rigorous experiments, we demonstrate that our method highly surpasses the baselines in terms of synthesized video quality both qualitatively and quantitatively[†].

1. Introduction

The task of video virtual try-on aims to synthesize a coherent video that preserves the appearance of one target clothes and the original person's pose and body shape in the source video. This task has attracted much attention in recent years because of the prospects of its wide application in e-commerce and short video industry.

Figure 1. Examples of our video virtual try-on results on the VVT dataset and our dataset (first row).

Previous virtual try-on methods usually focus on image-based operations [8, 12, 17, 18, 22, 28, 36, 44]. Among them, CP-VTON [36] proposed a geometric matching module to learn the parameters of TPS transformation, which greatly improves the accuracy of deformation. WUTON [22] and PFAFN [13] proposed parser-free methods to reduce the dependency of using accurate masks, and VITON-HD [8] further increased the resolution of the generated image. These methods have achieved great success in different aspects. However, image virtual try-on is far from video virtual try-on in terms of immersion, and it often lead to inconsistent results between frames when image-based methods applied to videos.

There are a few attempts of designing video virtual try-on. FW-GAN [9] is the first proposed method, which introduces the optical flow prediction module proposed in Video2Video [38] to warp the past frames to synthesize coherent future frames. Similarly, FashionMirror [7] also predicts optical flow, however, it warps the past frames to future frames at the feature level instead of the pixel level, which enabled it can generates clothes in different views. MV-TON [48] further adds a memory refinement module to memorize the features of past frames. Although the above methods have made some progress in video spatio-temporal consistency, however, their generated frames are flickering and has a distance from achieving a

---

spatio-temporally smooth video. We argue that there are two stems: on one hand, the above approaches only pay attention to try-on module while ignoring that the inputs deformed by the warping module are inconsistent. On the other hand, the images are synthesized in a frame by frame manner, which suffers from inconsistent attention results along spatio-temporal dimensions and often leads to blurriness and temporal artifacts in videos. Besides, the above approach is difficult to meet the demands in practical scenarios due to the presence of occlusions (*e.g.*, hair, hands, bags) appearing in the clothing area. Lastly, the above methods are designed for simple datasets with pure background, like VVT [9]. Thus they cannot deal with complex environment, nor to be in harmony with the natural background.

To address the challenges mentioned above, we propose a novel video virtual try-on framework, called Cloth-Former. Firstly, inspired by VITON-HD [8], we introduce a clothing-agnostic person representation that eliminates clothing information thoroughly and preserves background and occlusion. Next, we novelly employ frame-level TPS-based warp method to predict and mask the occlusion region of the target clothes, then feed the processed target clothes to an appearance-flow-based methods to get a accurate and anti-occlusion dense flow mapping (appearance-flow) between the body regions and the clothing regions. Moreover, we enforce output to be spatio-temporally consistent in both warp module and try-on module. In warp module, we carry out two steps (ridge regression and optical flow correction) on appearance-flow sequence to produce the temporally smooth warped clothes sequence. In try-on module, we propose a **M**ulti-scale **P**atch-based **D**ual-stream **T**ransformer (**MPDT**) generator with multi-input and multi-output to synthesize the final realistic video based on the outputs from the previous stages, which simultaneously optimizes all the output frames in a single feed-forward process. Lastly, MPDT generator employs MPDT block to extract the clothes color and texture spatio-temporal features from warped clothing sequence and extract preserved person features and environmental information in agnostic-clothing sequence, which aims to generate clothes, non-target body(including arms, neck *et al.*) and fill the background in agnostic region. Then, the integration between the clothing features and background content features adopted in MPDT block to generate more harmonious results. To validate the performance of our framework, we collected a wild virtual try-on dataset with occlusion and complicated background for our research purpose. Our experiments demonstrate that ClothFormer significantly outperforms the existing methods in generating videos, both quantitatively and qualitatively.

Our contributions can be summarized as below:

- A novel warp module that combines the advantages of TPS-based methods and appearance-flow-based meth-

ods is designed to address the problem of inaccurate warp due to occlusions appear in clothing region.

- A tracking module based on ridge regression and optical flow correction are proposed to deforme a temporally smooth warped clothing sequence, which provides a prerequisite for the try-on module to generate coherent videos.

- The MPDT generator is designed carefully in the try-on module, which can extract and fuse clothing textures, person features and environment information to generate realistic try-on videos. To the best of our knowledge, this is the first time that transformer has been applied to the video virtual try-on.

2. Related Work

Video processing and generation The video processing and generation techniques summarized here including video inpainting [26,43], video instance segmentation [39], video semantic segmentation [27], video super resolution [20], video-to-video synthesis [37, 38] and video virtual try-on [7,9,48]. These works share the common process of using a temporally consistent video sequence as input and harnessing temporal information to produce temporally smooth videos. Inspired by previous works, FW-GAN [9], FashionMirror [7] and MV-TON [48] adopted various video processing methods to the task of virtual try-on and have proved their effectiveness. FW-GAN and FashionMirror predict optical flow to warp the past synthesized frames at the pixel or feature level to generate subsequent frames, which was first proposed in vid2vid [38]. MV-TON proposes a memory refinement module to reconstruct spatio-temporal information, which was used in video semantic segmentation [27, 30]. However, these methods ignore that the input of try-on module, *i.e.*, the warped clothing sequence is not smooth in the temporal dimension, which leads to blurriness and temporal inconsistency in videos. In contrast, we propose a tracking strategy based on optical flow and ridge regression to obtain a temporally consistent warp sequence as the input of the try-on module.

Vision Transformer The transformer [35] was first proposed for sequence-to-sequence machine translation task and have been recently adopted for computer vision tasks. The ViT [10] was proposed by directly applying a pure transformer to sequences of 16x16 image patches for image classification tasks, which attains promising results compared to previous convolutional neural networks (CNNs). DETR [3] built an end-to-end object detection method using bipartite-matching loss with transformer-based encoder and decoder, which largely simplifies the traditional detection pipeline [32] and achieves high performance and efficiency on par with CNN-based methods. Inspired by DETR, VisTR [40] builts a simpler and faster video instance

segmentation framework by using transformer. STTN [46] proposed a multi-scale patch-based joint Spatial-Temporal Transformer Network (STTN) for video inpainting and achieved state-of-the-art performance. These methods have proved the effectiveness of transformer in computer vision tasks. However, to the best of our knowledge, there is no previous study that successfully applies transformer in video virtual try-on. We find that the transformer can not only explore spatial correlation between patches by self-attention mechanisms but can also extract temporal correlation across multiple frames. Based on this idea, we propose a Dual-Stream Transformer for video virtual try-on.

Virtual Try-on Existing methods on virtual try-on can be classified as 3D-model-based approaches [1, 16, 23, 29, 31] and 2D-image-based ones [8, 12, 17, 22, 28, 36, 44, 45]. 3D-model-based approaches are not widely applicable due to the need of additional 3D measurements and high computational complexity, while 2D-image-based approaches are more broadly applicable. VITON [18] designs a coarse-to-fine strategy which can seamlessly transfer a desired clothing item onto the corresponding region. The geometric matching module in CP-VTON [36] well preserves the clothes identity in the generated image. WUTON [22] and PFAFN [13] proposed parser-free methods to relieve the need of using accurate masks. VITON-HD [8] synthesizes 1024×768 images by using ALIgnment Aware Segment (ALIAS) normalization and ALIAS generator. Compared with image virtual try-on, video virtual try-on is more practical and user-friendly. FW-GAN [9] proposed a flow-navigated warping GAN in video virtual try-on to generate coherent video stream. MV-TON [48] adopted memory refinement to improve the details by modeling the previously generated frames. Nevertheless, none of the above video virtual try-on methods can deal with the temporal consistency of warped input sequences. In this paper, we propose to use dual-stream transformer to process warped input sequences, by temporally aggregating and smoothing background and foreground information in spatio-temporal dimension to synthesize realistic video stream.

3. Proposed Method

As described in Fig. 2, given a target clothing image $C \in \mathbb{R}^{3 \times H \times W}$ and a reference person video sequence $I_1^T := \{I_1, ..., I_T\} \in \mathbb{R}^{3 \times H \times W}$, H and W denote height and width of the image, T is the frame length of the sequence. ClothFormer aims to synthesis a realistic video sequence $\tilde{I}_1^T := \{\tilde{I}_1, ..., \tilde{I}_T\} \in \mathbb{R}^{3 \times H \times W}$ that represents a person wearing the target clothes C with spatio-temporal consistency, where the pose and body shape of I_1^T as well as the color and texture of C are preserved. Training with sample triplets $(I_1^T, C, \tilde{I}_1^T)$ is straightforward but undesirable in practice [36]. Instead, we use (I_1^T, C, I_1^T) where the clothes C already worn on the reference person video

sequence I_1^T.

Since directly training on (I_1^T, C, I_1^T) harms model's generalization ability during inference, we construct an occlusion preserved clothing-agnostic person representation that eliminates the effects of source clothing in I_1^T clarified in Sec. 3.1. We combine the advantages of the TPS-based warping methods and appearance-flow-based methods to solve the problem of inaccurate warping due to occlusion, and add a tracking module to deform a temporally smooth warped clothing sequence, which are introduced in detail in Sec. 3.2 and Sec. 3.3. Lastly, we propose a novel MPDT generator to synthesis realistic video in Sec. 3.4.

3.1. Pre-processing

Inspired by [8, 44], we propose constructing a clothing-agnostic video sequence A_1^T as inputs of warping and try-on modules, which preserves the person identity (*e.g.*, face, hands and lower body) and eliminates clothing-gnostic regions with the occlusion preserved. We obtain A_1^T by using the following four sequences, the segmentation map sequence $S_1^T := \{S_1, ..., S_T\}$, the Dense-Pose sequence $D_1^T := \{D_1, ..., D_T\}$, the pose sequence $P_1^T := \{P_1, ..., P_T\}$, and the matting sequence $M_1^T := \{M_0, M_1, ..., M_T\}$. We use the pre-trained networks [2, 14, 15, 25] to generate these sequences. Specifically, as shown in Fig. 2 (a), the A_1^T is derived through masking out the clothing-gnostic regions of the I_1^T by utilizing clothing-gnostic mask M_{a1}^T, where M_{a1}^T is expanded from arms, clothes and torsor-skin region predicted in S_1^T and then removing the hands and occlusion regions. Hands regions are predicted in D_1^T, the occlusion region is defined as the intersection of the foreground in M_1^T and the region predicted to be label zero in S_1^T, where label zero in S_1^T denotes background or other items such as backpack strap on I_1^T in Fig. 2 (a). Finally, the D_1^T, P_1^T and A_1^T are the inputs of warping module and try-on module to generate temporally consistent video $\tilde{I}_1^T$.

3.2. Frame-level Anti-occlusion Warp Module

The existing warping methods cannot handle occlusions that appear in clothing regions. Most of them utilize appearance-flow-based methods or TPS-based methods to deform clothing images. The appearance-flow-based methods are very sensitive to occlusions, when there are occlusions (*e.g.*, hair, arms, bags), the pixel-squeeze phenomenon is likely to occur as shown in the third column of Fig. 6. TPS-based methods can handle partial occlusions by estimating grid mapping to deform clothing images. However, it often leads to the misalignment between the warped clothes and the body [8]. To address these issues, we propose a two-stage anti-occlusion strategy.

In the first stage, we adopt a TPS-based warping module. As shown in Fig. 2(b), we use (A_t, D_t, P_t) and the reference
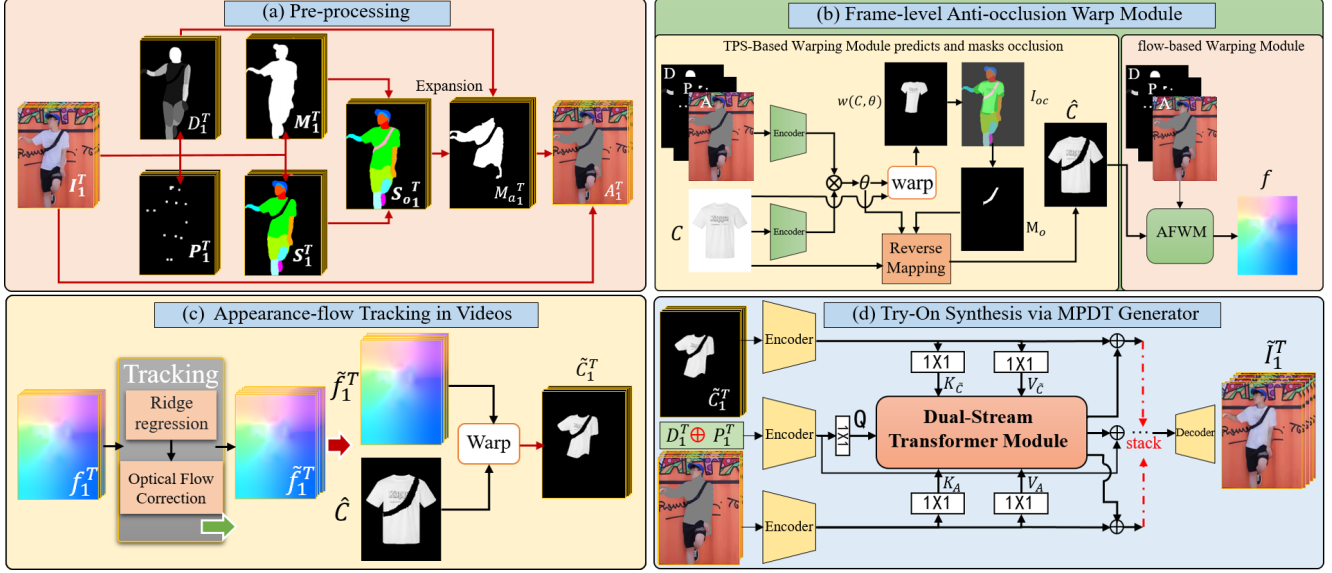
Figure 2. Framework of ClothFormer. (a) First, we obtain clothing-agnostic person image sequences A_1^T. (b) We predict t warped clothes $W(C, \theta_t)$ by TPS-based warp method to infer an anti-occlusion target clothes C, then appearance-flow-based warp method is adopted to get an appearance flow f. (c) Appearance-flow tracking module based on ridge regression and optical flow correction is designed to get warped clothing sequence with spatio-temporally consistent. (d) Finally, MPDT generator synthesizes the final output video sequence $\tilde{I}_1^T$ based on the outputs from the previous stages

clothes C as inputs at frame t. The TPS-based warping loss is formulated as:

$$L_t^{TPS\text{-warp}} = \left\| I_t^C - W(C, \theta_t) \right\|_1 + \lambda_t^{sdc} L_t^{sdc} \qquad (1)$$

where I_t^C is the target clothes extracted from I_t at frame t, $W(C, \theta_t)$ is the warped clothes that deforms C using θ_t, and λ_t^{sdc} is the hyper-parameter for second-order difference constraint L_t^{sdc} [44].

Afterward, we define the region where $W(C, \theta_t)$ overlaps with the location of occlusion in S_{ot} as occluded region of $W(C, \theta_t)$, as I_{oc} shown in Fig. 2(b), and the occluded area shown as M_o. Then we get the target clothes $\hat{C}_t$ that have masked the occluded area by using θ_t to reverse mapping. Finally, AFWM [13] network adopted to learn dense flow mapping between the body regions and the clothing regions (appearance-flow f_t) with (A_t, D_t, P_t) and the $\hat{C}_t$ as inputs. The appearance-flow f_t is optimized as follows:

$$L_t^{flow-\text{warp}} = \left\| I_t^C - W(\hat{C}_t, f_t) \right\| + \lambda_t^{sec} L_t^{sec} \qquad (2)$$

where $W(\hat{C}_t, f_t)$ is the warped clothes that deforms $\hat{C}_t$ using learned appearance-flow f_t at frame t, and λ_t^{sec} is the hyper-parameter for second-order smooth constraint L_t^{sec} [13].

3.3. Appearance-flow Tracking in Videos

For video virtual try-on, the texture and the color of the synthesized clothes are mainly related to the input warped clothing sequence. Previous works [7, 9, 48] only focused on the temporal consistency of the try-on module but ignored the temporal consistency of the warped clothing sequence. Instead, we propose a appearance-flow tracking module to produce a temporally smooth warped clothing sequence by tracking the appearance-flow learned in Sec. 3.2 as shown in Fig. 2(c). The appearance-flow learned in warping module represents the coordinates of the pixels in the input clothes $\hat{C}_t$ mapped to which position in the warped clothes $W(\hat{C}_t, f_t)$, from this point of view the appearance-flow-based warp module is similar to facial landmark detection task [42], the $\hat{C}_t$ and $W(\hat{C}_t, f_t)$ are analogous to aligned face image and the face image with different poses in facial landmark detection task. Inspired by [33] [11], we first reshape the f_t with the height H and width W into 1-dimension vectors f_t^{1D} of length $W \times H$, and track it by developing a post-processing algorithm based on ridge regression which exploits correlation among adjacent flow to achieve a temporally smooth results. The appearance-flow optimized as:

$$\hat{f}_t^{1D} = X \left(X^T X - \mu I \right) X^T f_t^{1D} \qquad (3)$$

where f_t is the appearance-flow of frame t and $1 \leq t \leq T$, I is the unit matrix, μ is the hyper parameter and X is the feature matrix detailed in the supplementary.

In addition, the motion information in the clothing area of input person sequence I_1^T is crucial because it is not only related to the human pose but also related to the environmental factors such as wind. Therefore, we used the optical

flow [21] in the clothing area to correct the $\hat{f}_t$, denoted as:

$$\tilde{f}_t = \begin{cases} \dfrac{\hat{f}_t + \widetilde{w}_{t-1}\left(\hat{f}_{t-1}\right)}{2}, & \delta_t \leq \varepsilon \quad and \quad \hat{f}_t \in \Omega \\ \tilde{f}_t = \hat{f}_t \quad, & \delta_t > \varepsilon \quad or \quad \hat{f}_t \notin \Omega \end{cases} \quad (4)$$

where $\hat{f}_t$ is the tracking result of Eq. (3), $\widetilde{w}_{t-1}$ is the estimated optical flow from I_{t-1} to I_t of input person sequence. Ω denotes the intersection of the clothing region on I_t and the warped clothing region of $W(\hat{C}_t, \hat{f}_t)$. By $\widetilde{w}_{t-1}(\hat{f}_{t-1})$, we warp $\hat{f}_{t-1}$ based on $\widetilde{w}_{t-1}$. δ_t defined as $\left\| \hat{f}_t - \widetilde{w}_{t-1}\left(\hat{f}_{t-1}\right) \right\|$ and we set threshold value ε to 0.05.

Finally, the warped clothing sequence with temporal consistency and anti-occlusion obtained as :

$$\tilde{C}_1^T = W(\hat{C}_1^T, \tilde{f}_1^T) \quad (5)$$

3.4. Try-On Synthesis via MPDT Generator

In the try-on module, we propose the MPDT generator to synthesize realistic video sequences based on the outputs of the previous stages. MPDT generator can deal with the blurriness and temporal artifacts from which previous methods [9, 48] suffer a lot.

As shown in Fig. 3, there are three inputs for MPDT generator: (1) warped clothing sequence $\tilde{C}_1^T$ generated in Sec. 3.3; (2) person-shape sequence $D_1^T \oplus P_1^T$ concatenated by DensePose sequence D_1^T and the pose sequence P_1^T generated in Sec. 3.1; (3) the clothing-agnostic sequence A_1^T. For model architecture, MPDT generator consists of three components: three frame-level encoders, MPDT block, and a frame-level decoder. The MPDT block based on Transformer is the core component which aims to aggregate spatio-temporal features from the warped clothing stream and the clothing-agnostic stream.

Embedding The embedding of query and key-value pairs play a crucial role in Transformer. For MPDT block, there are two sets of key-value pairs embedding are designed and a set of queries:

$$\begin{aligned} q_t &= \text{conv}_q\left(p_t\right) \\ \left(k_t^C, v_t^C\right) &= \left(\text{conv}_{k^C}\left(C_t\right), \text{conv}_{v^C}\left(C_t\right)\right) \\ \left(k_t^A, v_t^A\right) &= \left(\text{conv}_{k^A}\left(A_t\right), \text{conv}_{v^A}\left(A_t\right)\right) \end{aligned} \quad (6)$$

where $1 \leq t \leq T$, conv denotes the 1x1 2D convolutions.

Dual-stream Spatio-temporal Attention Inspired by STTN [46], we conduct attention among all spatio-temporal patches in the input video sequence. Each frame is segmented into patches of size $r_1 \times r_2 \times cn$ with the channel size cn, thus there are $N = T \times h/r_1 \times w/r_2$ patches for the entire sequence totally. Before calculating patch-wise similarities, the query and key patches are reshaped into 1D vectors. The attention weights between query patches and clothing key-value patches denote as:

$$\alpha_{i,j}^C = \begin{cases} \text{softmax}_j\left(\dfrac{p_i^q \cdot \left(p_j^{k^C}\right)^T}{\sqrt{r_1 \times r_2 \times cn}} \right), & p_j^{k^C} \in \Omega^C \\ 0, & p_j^{k^C} \notin \Omega^C \end{cases} \quad (7)$$

$$Att_i^C = \sum_{j=1}^{N} \alpha_{i,j}^C \cdot p_j^{v^C} \quad (8)$$

where p_i^q denotes i^{th} query patch. $p_j^{k^C}$ and $p_j^{v^C}$ denote the jth key-value patches. Ω^C denotes the visible clothing regions. Similarly, the attention value Att_i^A between query patches and clothing-agnostic patches is calculated in the same way as Eq. (8). Then attention is applied in a multi-head manner and the attention values from different heads are concatenated as Att^C and Att^A. We fuse these two streams through concatenation followed by a 1x1 convolution:

$$o = \left(Att^C \oplus Att^A\right) W_1 + b_1 \quad (9)$$

where $\oplus$ denotes the concatenation, W_1 and b_1 are learnable parameters of 1x1 convolutions.

The result o and query q (added by residual connection) in current MPDT block will serve as the query of the next block. Then, a frame-level decoder simultaneously renders person image sequence I_{R1}^T and predicts composition mask sequence M_{C1}^T. We fuse I_{R1}^T and the warped clothing sequence $\tilde{C}_1^T$ by using M_{C1}^T to enhance the texture details of the generated clothes, it is essential for $\tilde{C}_1^T$ to be temporally smooth.

$$I_{\text{masked}1}^T = M_{C1}^T \odot \tilde{C}_1^T + (1 - M_{C1}^T) \odot I_{R1}^T \quad (10)$$

To reconstruct complex background and focus on the task of virtual try-on, the clothing-agnostic image and $I_{\text{masked}1}^T$ are then fused together using the clothing-gnostic mask sequence M_{a1}^T defined in Sec. 3.1 to synthesize the final output $\tilde{I}_1^T$:

$$\tilde{I}_1^T = (1 - M_{a1}^T) \odot I_{\text{masked}1}^T + M_{a1}^T \odot A_1^T \quad (11)$$

Finally, we use spatio-temporal losses to train MPDT. In spatial dimension, we include l_1 loss and perceptual loss [24] to ensure per-pixel reconstruction accuracy. In temporal dimension, we use a Temporal PatchGAN (TP-GAN) [5, 6] as the discriminator to improve the temporal consistency in generated video. The overall objective function is as:

$$L_{\text{try-on}} = \lambda_1 L_{l1}^{\text{whole}} + \lambda_2 L_{l1}^{clothes} + \lambda_3 L_{\text{perc}} + \lambda_4 L_{\text{TPGAN}} \quad (12)$$

where L_{l1}^{whole} denotes the L1 loss of the whole image, $L_{l1}^{clothes}$ denotes the L1 loss in clothing regions. L_{TPGAN} is the adversarial loss. $\lambda_i, i \in \{1, 2, 3, 4\}$ are hyper-parameters.
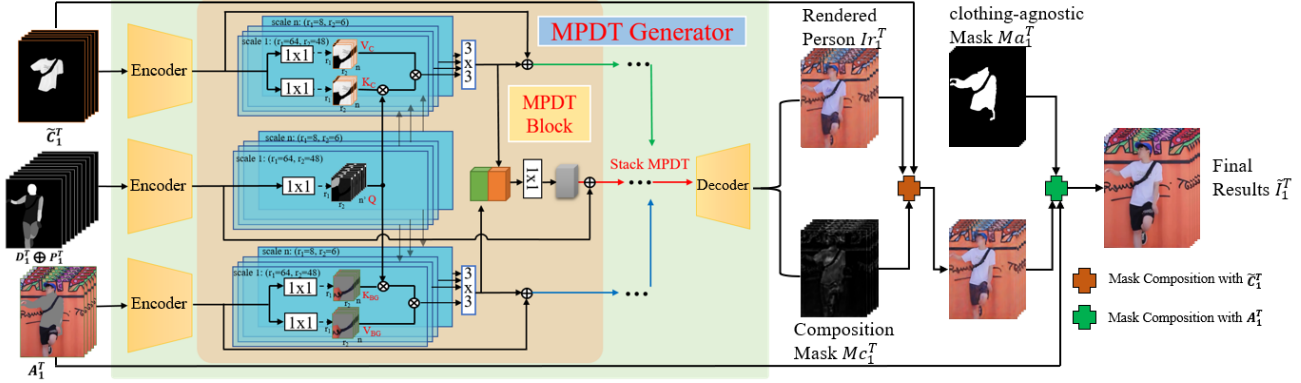
Figure 3. Illustration of the MPDT generator, including three frame-level encoders, stacked **MPDT** block, and a frame-level decoder. MPDT block is the core module, on one hand, MPDT block search and extract the texture and the color content from warped clothing sequence $\tilde{C}_1^T$, on the other hand MPDT block borrow the environment and person identity information from clothing-agnostic person image sequence A_1^T to synthesis body, fill masked background and make the generated clothes more harmonious with the environment.

Method	Dataset	SSIM	LPIPS	VFID I3D	VFID ResNeXt101
CP-VTON [36]	VVT	0.459	0.535	6.361	12.10
ACGPN [44]	VVT	0.853	0.178	9.777	11.98
PBAFN [13]	VVT	0.870	0.157	4.516	8.690
FW-GAN [9]	VVT	0.675	0.283	8.019	12.15
MVTON [48]	VVT	0.853	0.233	8.367	9.702
ClothFormer	VVT	**0.921**	**0.081**	**3.967**	**5.048**
CP-VTON [36]	ours	0.682	0.299	13.11	31.19
ACGPN [44]	ours	0.786	0.243	16.21	32.54
PBAFN [13]	ours	0.841	0.188	11.15	28.62
FW-GAN [13]	ours	0.705	0.344	13.71	28.31
CP-VTON* [36]	ours	0.929	0.068	7.463	11.30
ACGPN* [44]	ours	0.936	0.066	10.89	13.91
PBAFN* [13]	ours	0.932	0.066	6.132	10.88
ClothFormer†	ours	0.953	0.047	5.071	9.018
ClothFormer✳	ours	**0.959**	0.042	5.140	9.394
ClothFormer◇	ours	0.949	0.050	5.653	9.721
ClothFormer-tiny	ours	0.955	0.045	5.208	9.153
ClothFormer	ours	0.958	**0.042**	**5.024**	**8.971**

Table 1. Comparison with previous methods on the VVT dataset and our new collected dataset. For SSIM, the higher is the better. For LPIPS and VFID, the lower is the better. ClothFormer†, ClothFormer✳, ClothFormer◇ and ClothFormer-tiny are Cloth-Former variants for ablation study.

4. Experiments

4.1. Experiment Setup

Datasets. Experiments are conducted on the VVT [9] dataset and our collected dataset. The VVT dataset contains 791 videos with the resolution of 192×256. The train and test set contain 159,170 and 30,931 frames respectively.

However, in the VVT dataset, the backgrounds are primarily white, and the human poses are monotonous and simple. In contrast, we collected a wild virtual try-on dataset with complex environment, complicated poses and occlusions, which contains 3995 videos. The train and test set contain 1799,65 and 25,710 frames respectively. In addition, every video in our dataset is divided into several coherent sub videos by adopting shot transition detection [34].

Training and Testing. We train the warp module and the try-on module separately and combine them to generate the try-on image eventually. The paired setting used to train the modules in the training process. In the testing process, we use the pairs of a person and a clothing image to evaluate a paired setting and shuffle the clothing images for an unpaired setting as in previous methods [8] [44]. Moreover, the videos in VVT dataset are divided into coherent sub videos to train and test methods.

4.2. Qualitative Analysis

We first compare proposed methods with video-based method FW-GAN [9], MV-TON [48] and image-based methods CP-VTON [36], ACGPN [44] and PB-AFN [13] in VVT dataset for a more comprehensive experimental comparison. Moreover, to verify the superior performance of our model in complex environments and with occlusions appear in the person images, ClothFormer compared with all the above methods except MVTON [48] in our dataset for whose testing code and pretrained model on VVT dataset are available while the training code is not. (**Notes:** It is necessary to watch the videos to compare the qualitative results at urls or supplementary materials).

Fig. 4 shows some qualitative results on the VVT dataset. Clothes generated by CP-VTON, ACGPN, FW-GAN, MV-TON show many visual artifacts, including blurring and cluttered texture. Although each frame synthesized by PB-AFN is photorealistic, the resulting video lacks temporal
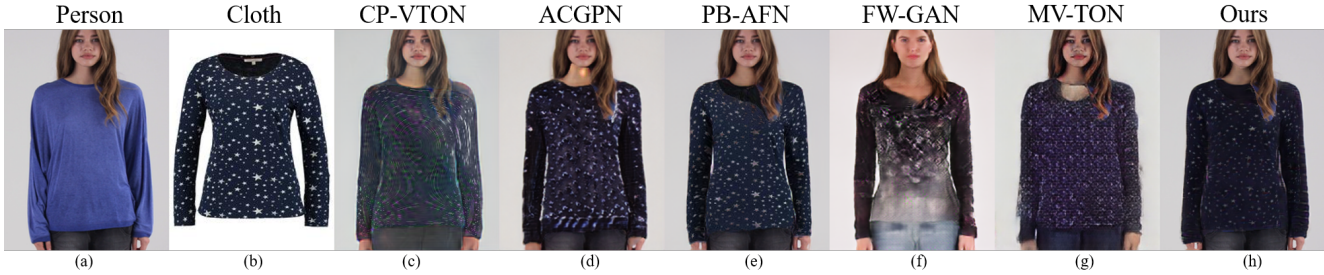
Figure 4. Visual comparison with the baseline methods on the VVT dataset. (a) is the reference person, (b) is the target clothes. Cloth-Former produces a more temporally consistent video output and clearly preserve the details of the target clothes.

Figure 5. Visual compare with the baseline methods on our dataset, (b) is the clothing-agnostic images for composition as Eq. (11) when training CP-VTON*, ACGPN*, PB-AFN* and ClothFormer. The first row shows ClothFormer generates a more harmonious results, the second row shows ClothFormer generates more satisfactory results when cross-arms appear in clothing region.

coherence. Specifically, the texture of the clothes generated by PB-AFN is flickering irregularly even when the person keeps still. Compared with the baseline methods, Cloth-Fomer produces a temporally consistent video output and preserves the details of the target clothes.

To compare the performance of ClothFormer and the baseline methods in complex environments and the person with occlusions appearing in the clothing region, we conduct experiments in our new collected dataset. We first trained baseline methods CP-VTON, ACGPN and PBAFN according to their original setting. However, unlike the VVT dataset, the baseline methods could reconstruct the white background; the background of the reference person in our dataset is too complicated to reconstruct for these methods. Hence, for a fair comparison, like ClothFormer, we fuse the clothing-agnostic region and the output of these method during training process, which are denoted as CP-VTON*, ACGPN* and PBAFN* respectively. From the sample results in the first row of Fig. 5, it can be easily observed that our method achieves much better visual consistency compared to the baseline methods. For one thing, the clothes and body generated by baseline methods look unnat-

ural in complex environments while the clothes and body generated by ClothFomer are in perfect harmony with the complex background. For another thing, ClothFormer can fill the region around generated body and clothes with plausible content in a video while the baseline methods failed. From the sample results in the second row of Fig. 5, when occlusions like cross-arms appear in the person images, PB-AFN generates unnatural results with pixel-squeeze phenomenon, CP-VTON fails to generate arms and ACGPN generates fake and flickering arms. In comparison, our ClothFormer can warp the clothes to the target person accurately even when occlusion appears. The two rows of FW-GAN results do not even generate a humanoid , since FW-GAN only uses the RGB information of the first frame and the pose information of the subsequent frames to generate video, it inevitably achieves unsatisfying performance training in our dateset with complicated background and complicated poses.

4.3. Quantitative Analysis

As shown in Tab. 1, we perform the quantitative experiments in terms of both image-based evaluation metrics and
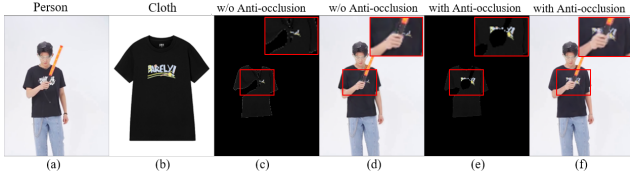
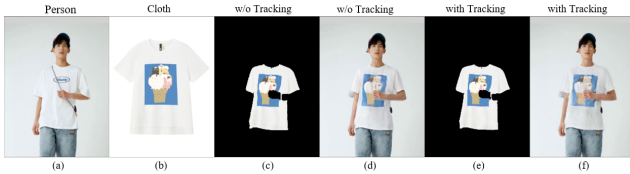Figure 6. Effects of Anti-occlusion Warp Module.

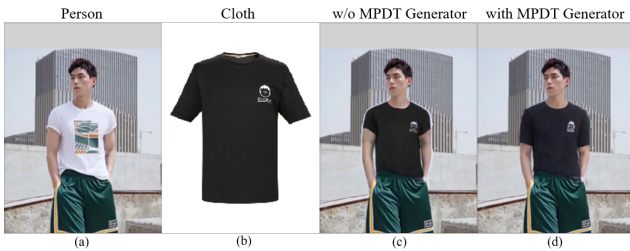Figure 7. Effects of the Appearance-flow Tracking Module.

Figure 8. Effects of MPDT generator.

video-based evaluation metrics. For image results, We use the structural similarity (SSIM) [41] and the learned perceptual image patch similarity (LPIPS) [47] to evaluate our method in the paired setting. For video results, we use the Video Frechet Inception Distance (VFID) to measure visual quality and temporal consistency in the unpaired setting, and both temporal and spatial features are extracted by two pre-trained video recognition CNN backbones: I3D [4] and 3D-ResNeXt101 [19].

Compared with image-based methods and video-based methods on VVT dataset, ClothFormer outperforms them by a large margin, ClothFormer also surpasses the baseline methods with or without fusing the clothing-agnostic region on our new collected dataset, which demonstrates Cloth-Former has great advantage in generating high-quality and spatio-temporally consistent try-on videos.

In pursuit of speed improvement, we try to compress ClothFormer by reducing the number of channels of MPDT from 256 to 96 and stacks of blocks from 8 to 6(ClothFormer-tiny), the FLOPs are reduced from 70.29G to 10.63G. However, the quantitative metrics are still better than other methods.

4.4. Ablation Study

We conduct an ablation study to analyze the designed modules of our method by creating three variants in our collected dataset, the Anti-occlusion Warp Module, Appearance-flow Tracking module and MPDT generator.

Effectiveness of Anti-occlusion Warp Module As

shown in Fig. 6, the ClothFormer† without Anti-occlusion Warp Module generates artifacts around occlusion with Pixel-squeeze phenomenon in the red box region of both warped clothes sequence and generated results. In contrast, our ClothFormer could generate satisfactory results, which demonstrates that the Anti-occlusion Warp Module has the capability to generate accurate warping against occlusion.

Effectiveness of Appearance-flow Tracking Fig. 7 shows the texture of the clothes generated by ClothFormer✳ without Appearance-flow Tracking Module is flickering irregularly, verify Appearance-flow Tracking is beneficial to module produces a more temporally smooth result.

Effectiveness of MPDT Generator As shown in Fig. 8, the result of ClothFormer◇ with a U-Net try-on generator shows some weaknesses, such as, the non-target body parts looks fake, the region around the upper-body are blurring and the generated clothes are not harmony with the background. On the contrary, our ClothFormer with MPDT generator is capable to synthesize more video-realistic, natural and pleasing results, which demonstrates the superiority of dual-stream structure in the MPDT block.

In addition, as shown in Tab. 1, the VFID scores of ClothFormer† are close to ClothFormer while there are gaps in SSIM and LPIPS, on the contrary, the values of SSIM and LPIPS of ClothFormer✳ are almost identical with Colth-Former while ClothFormer outperforms on VFID, which demonstrates that the Anti-occlusion Warp Module is beneficial to generate more accurate warping result to synthesize more photo-realistic images while the Appearance-flow Tracking Module helps to generate more temporally smooth videos. And Tab. 1 also shows ClothFormer outperforms ClothFormer◇ on all indexes, which demonstrates the effectiveness of our MPDT generator.

5. Conclusions

We propose a novel video virtual try-on framework, *i.e.* ClothFormer, which aims at generating realistic try-on videos while preserving the character of clothes, details of human identity (posture, body parts, bottom clothes) and background. We present three carefully designed modules, *i.e.* Frame-level Anti-occlusion Warp module, Appearance-flow Tracking module and MPDT generator. Qualitative and quantitative experiments demonstrate that ClothFormer surpasses existing virtual try-on methods with a large margin.

Acknowledgements

References

[1] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5420–5430, 2019. 3

[2] 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. 3

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 8

[5] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9066–9075, 2019. 5

[6] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Learnable gated temporal shift module for deep video inpainting. *arXiv preprint arXiv:1907.01131*, 2019. 5

[7] Chieh-Yun Chen, Ling Lo, Pin-Jui Huang, Hong-Han Shuai, and Wen-Huang Cheng. Fashionmirror: Co-attention feature-remapping virtual try-on with sequential template poses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13809–13818, 2021. 1, 2, 4

[8] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14131–14140, 2021. 1, 2, 3, 6

[9] 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/CVF International Conference on Computer Vision*, pages 1161–1170, 2019. 1, 2, 3, 4, 5, 6

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[11] Zhenye Gan, Lizhuang Ma, Chengjie Wang, and Yicong Liang. Improved cnn-based facial landmarks tracking via ridge regression at 150 fps on mobile devices. In *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–9. IEEE, 2017. 4

[12] Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, and Ping Luo. Disentangled cycle consistency for highly-realistic virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16928–16937, 2021. 1, 3

[13] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8485–8493, 2021. 1, 3, 4, 6

[14] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 770–785, 2018. 3

[15] 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

[16] Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. Garnet: A two-stream network for fast and accurate 3d cloth draping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8739–8748, 2019. 3

[17] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10471–10480, 2019. 1, 3

[18] 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, 3

[19] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 8

[20] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3897–3906, 2019. 2

[21] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 5

[22] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzenes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 619–635. Springer, 2020. 1, 3

[23] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision*, pages 18–35. Springer, 2020. 3

[24] 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

[25] Zhanghan Ke, Kaican Li, Yurou Zhou, Qiuhua Wu, Xiangyu Mao, Qiong Yan, and Rynson WH Lau. Is a green screen really necessary for real-time portrait matting? *arXiv preprint arXiv:2011.11961*, 2020. 3

[26] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5792–5801, 2019. 2

[27] Jiaxu Miao, Yunchao Wei, and Yi Yang. Memory aggregation networks for efficient interactive video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10366–10375, 2020. 2

[28] 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 *CVPR Workshops*, 2020. 1, 3

[29] Aymen Mir, Thiemo Alldieck, and Gerard Pons-Moll. Learning to transfer texture from clothing images to 3d humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7023–7034, 2020. 3

[30] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019. 2

[31] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7365–7375, 2020. 3

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 2

[33] Enrique Sánchez-Lozano, Georgios Tzimiropoulos, Brais Martinez, Fernando De la Torre, and Michel Valstar. A functional regression approach to facial landmark tracking. *IEEE transactions on pattern analysis and machine intelligence*, 40(9):2037–2050, 2017. 4

[34] Tomáš Souček and Jakub Lokoč. Transnet v2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020. 6

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[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, 3, 6

[37] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. *arXiv preprint arXiv:1910.12713*, 2019. 2

[38] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018. 1, 2

[39] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2021. 2

[40] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2021. 2

[41] 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

[42] Yue Wu and Qiang Ji. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127(2):115–142, 2019. 4

[43] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2019. 2

[44] 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, 3, 4, 6

[45] 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/CVF International Conference on Computer Vision*, pages 10511–10520, 2019. 3

[46] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *European Conference on Computer Vision*, pages 528–543. Springer, 2020. 3, 5

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 8

[48] Xiaojing Zhong, Zhonghua Wu, Taizhe Tan, Guosheng Lin, and Qingyao Wu. Mv-ton: Memory-based video virtual try-on network. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 908–916, 2021. 1, 2, 3, 4, 5, 6