Unsupervised Image-to-Video Clothing Transfer

A. Pumarola[1] V. Goswami[2] F. Vicente[3] F. De la Torre[3] F. Moreno-Noguer[1]
[1]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain
[2]Facebook AI Research, Menlo Park, USA
[3]Facebook, Menlo Park, USA

Figure 1: **Example of visual clothing transferring.** Left, original image of an iconic computer science clothing style. Right, the result of transferring the cloths from the left image to other images containing one or several subjects in unconstrained poses and background. This figure illustrates results in individual images, but the system is able generate space-time consistent novel views of clothing in videos.

Abstract

We present a system to photo-realistically transfer the clothing of a person in a reference image into another person in an unconstrained image or video. Our architecture is based on a GAN equipped with a physical memory that updates an initially incomplete texture map of the clothes that is progressively completed with the new inferred occluded parts. The system is trained in an unsupervised manner. The results are visually appealing and open the possibility to be used in the future as a quick virtual try-on clothing system.

1 Introduction

Virtual dressing rooms are expected to have a major impact in the fashion e-commerce industry. A major limitation of existing systems is that they rely on expensive setups (e.g. depth cameras) and/or require building sophisticated physical models of the clothing. We present a simple yet effective solution to the problem, which does not require modeling the underlying physics of the clothes, while still producing photo-realistic results. Fig. 1 illustrates the problem that this paper addresses. Our model is able to synthesize space-time consistent novel views of the source clothing, while simultaneously fitting them to the target person body shape and maintaining the original background. The proposed method is learned in an unsupervised fashion, that is, we do not require pairs of images of the same person with same clothes in different positions.

To address all these challenges, we combined a clothing segmentation output with a temporally-consistent Generative Adversarial Network (GAN). Our main contribution consists in equipping a standard GAN architecture with a memory module that progressively refines a source texture map and adapts it to the target person, by filling occluded regions and adapting to new lighting conditions and body pose. This work is related to recently proposed deep-learning approaches for transferring clothes [4, 10]; however, while these models provide visually compelling results, they typically rely on 3D human models, and their results are limited to non-cluttered backgrounds, mild lighting conditions and require supervised training. Our approach offers a simple but effective *unsupervised* image2Video approach that is shown to be robust results across pose, background, lighting and body variability without the need of knowing the underlying geometry of the body nor the physics ruling the cloth deformations.

2 Problem Formulation

Let $\mathbf{I}_c \in \mathbb{R}^{H\times W\times 3}$ be an input RGB image of a dressed person (source), and let $\mathbf{x}_1^T = (\mathbf{X}_1, \dots, \mathbf{X}_T)$ be the target video, where $\mathbf{X}_t \in \mathbb{R}^{H\times W\times 3}$ and the subindex t denotes the video frame. The target video can be of the same person in $\mathbf{I}_c$ or a different one. Our goal is to learn a mapping $\mathcal{M}$ to transform $\mathbf{x}_1^T$ into an output video $\mathbf{y}_1^T$ where

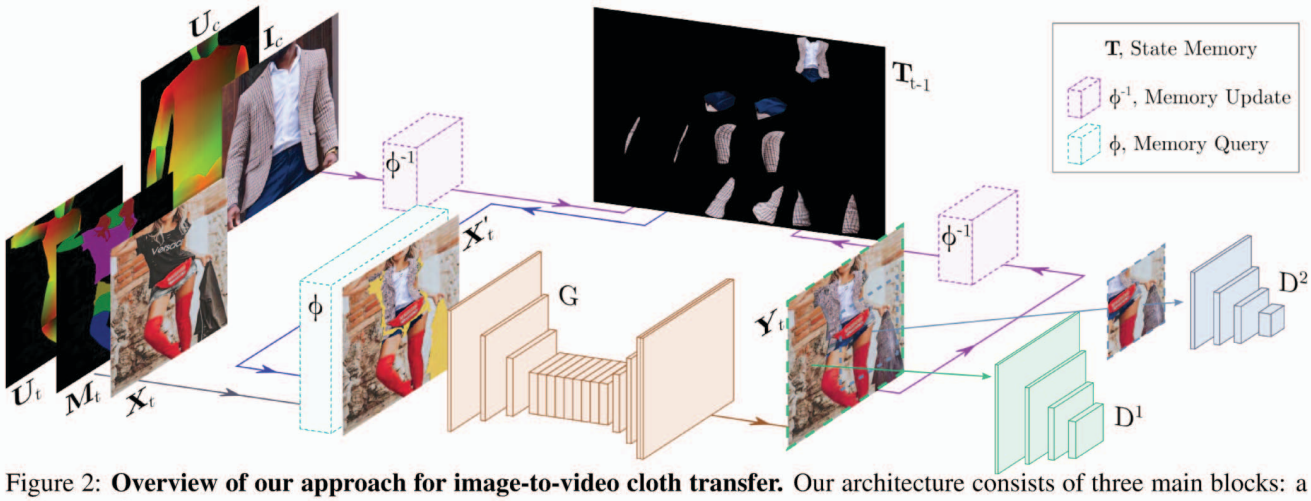

Figure 2: **Overview of our approach for image-to-video cloth transfer.** Our architecture consists of three main blocks: a generator G to transfer cloth items; a memory $\mathbf{T}$ that stores textures across long video sequences; and a multiscale discriminator D to evaluate the photo-realism of the generated image and its consistency with the cloth segmentation labels.

target person is realistically dressed with the clothes of $\mathbf{I}_c$. That is, we aim to transfer cloth items from the source image $\mathbf{I}_c$ to the target video $\mathbf{x}_1^T$ by learning the mapping $\mathcal{M} : (\mathbf{x}_1^T, \mathbf{I}_c) \rightarrow \mathbf{y}_1^T = (\mathbf{Y}_1, \ldots, \mathbf{Y}_T)$. One of our major contributions is to learn $\mathcal{M}$ in an unsupervised manner, that is, **we do not require pairs of images of the same person under different clothes or poses, or the same person wearing the same clothes under varying body postures**. Instead, our training data consists merely of N input RGB images $\{\mathbf{I}_c^n\}_{n=1}^N$.

The output video $\mathbf{y}_1^T$ is expected to be meet the following criteria: (i) the transferred cloth items must be adjusted to the pose and body shape of the target person; (ii) hallucinated views of the source clothes that are not visible in the source image $\mathbf{I}_c$, must be photo-realistic and consistent with the visible parts; (iii) transferred cloth items must be consistent with the illumination of the target video $\mathbf{x}_1^T$; (iv) target elements as body parts, background and non-transferred cloth items must remain fixed; and (v) the texture across the output frames of $\mathbf{y}_1^T$ must be consistent in space and time.

Dataset Pre-processing: In order to learn the cloth mapping transformation, we automatically enrich the input dataset of RGB images with segmentation masks $\{\mathbf{M}^n\}_{n=1}^N$ computed for each input image $\mathbf{I}_c^n$, where $\mathbf{M}^n \in \mathbb{R}^{H \times W \times S}$ are the segmentation masks for S cloth labels. We further augment the dataset with random occlusions, rotations, translations and color jitter. Specifically, occlusions are randomly introduced over the cloth and body regions (excluding the face) to simulate non-visible parts of the source clothing to be hallucinated. To help improving the blending of the transferred cloth onto the image background we also add occlusions on the cloth-background boundaries.

Cloth Segmentation: Segmentation labels will be used to guide the cloth transfer process between people under potentially different outfits. Therefore, we define a reduced number of high-level categories that can be shared even under different clothing styles. Concretely, the segmentation labels we consider are: hair, skin, top-layer1, top-layer2, bottom and shoes. For estimating such segmentation labels we use a PSPNet architecture [11] with a Resnet50 backbone, initialized with pre-trained Imagenet weights.

3 Image-to-Video Cloth Transfer

We next describe the main components of our GAN to photo-realistically transfer clothing (Fig. 2).

Texture Memory: It corresponds to the estimated texture map of the target cloth. We use the same body partition and UV parametrization as in DensePose [1]. At $t = 0$ the memory is initialized with the cloth's visible parts from the source image $\mathbf{I}_c$. At each time step, the cloth regions not seen in the target image are hallucinated by the generator and cumulatively added to the state memory. An example of how the memory evolves across a video sequence is shown in the top of Figure 3.

Memory Query: The texture memory can be accessed by both the discriminator an the generator. In the generation phase, the memory is queried using the mapping $\Phi : (\mathbf{X}_t, \mathbf{U}_t, \mathbf{M}_t, \mathbf{T}_{t-1}) \rightarrow \mathbf{X}'_t$, which renders the source cloth into the target frame $\mathbf{X}_t$. In order to perform this mapping, we first extract dense 2D correspondences $\mathbf{U}_t$ between the input the image $\mathbf{X}_t$ and a 3D body model, which implicitly provides the mapping onto the texture map. The correspondences are obtained with the pretrained DensePose network [1]. This initial mapping is still incomplete, and $\mathbf{X}'_t$ contains several regions with missing information.

Cloth Segmentation: The segmentation mask $\mathbf{M}_k$ for each video frame is inferred using the network we described in Sect. 2. This network performs the mapping $\Omega : (\mathbf{X}_k, \mathbf{M}_{k-1}) \rightarrow \mathbf{M}_k$.

Figure 3: **Memory state and segmentation tracking. Top-Left:** Input source image $\mathbf{I}_c$. **Bottom-Left:** First frame $\mathbf{X}_t$ of the target video in which we seek to transfer the clothes of $\mathbf{I}_c$. **Top-row (columns 2-5)**: Visualization of the memory $\mathbf{T}_t$, initially containing only the parts visible in $\mathbf{I}_c$. Novel regions hallucinated for the frames $\mathbf{X}_t$ are progressively added into the texture map. **Bottom-row (columns 2-5)**: Segmentation masks $\mathbf{M}_t$ (automatically estimated) and cloth transfer results $\mathbf{Y}_t$.

Generator: The incomplete image $\mathbf{X}'_t$ and the input segmentation masks are passed to the generator $G : (\mathbf{X}'_t, \mathbf{M}_t) \rightarrow \mathbf{Y}_t$. We force G to primarily focus on the segmented regions of the body, by adapting the lighting in the regions of $\mathbf{X}'_t$ which already have texture information, and inpainting those which do not have.

Memory Update: The new regions in $\mathbf{Y}_t$ the generator has hallucinated are mapped back to the texture memory using the inverse of the mapping we considered during the memory query phase, that is, $\Phi^{-1} : (\mathbf{Y}_t, \mathbf{M}_t, \mathbf{U}_t) \rightarrow \mathbf{T}_t$.

Multilevel Discriminator: The photo-realism of the generated image $\mathbf{Y}_t$ is evaluated with the network $D(\mathbf{Y}_t, \mathbf{M}_t)$. Its structure is similar to the multilevel PatchGan [9], which is made of two discriminators with identical architecture that operate at different image resolutions, one having a global view of the image to guide the generator to produce cloth labels, and the other focused on fine texture details.

3.1 Learning the Model

We train our model with a loss made of three terms:

Image Adversarial Loss ($\mathcal{L}_I$): We extend the standard min-max strategy [3] to enforce the model not just to produce photo-realistic images but also to be consistent with the cloth segmentation labels. Concretely, we add an extra term in the adversarial loss that aims to classify a mismatched image-mask pair as a negative sample. Formally, let $\mathbf{X}$ be the input image with cloth segmentation labels $\mathbf{M}$; $\mathbb{P}_r$ the data distribution of the input images and $\mathbb{P}_g$ the distribution of the generated images $\hat{\mathbf{X}} = G(\mathbf{X}', \mathbf{M})$; and $\hat{\mathbf{M}}$ a segmentation mask randomly chosen from the training set. The extended adversarial loss $\mathcal{L}_I$ is defined as:

$$\mathcal{L}_I = \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_r}[\log(D(\mathbf{X}, \mathbf{M}))] + \lambda(\mathbb{E}_{\mathbf{X}\sim\mathbb{P}_r}[\log(1 - D(\mathbf{X}, \hat{\mathbf{M}}))] + \mathbb{E}_{\hat{\mathbf{X}}\sim\mathbb{P}_g}[\log(1 - D(\hat{\mathbf{X}}, \mathbf{M}))]), \quad (1)$$

where $\lambda = 0.5$ allows balancing the positive-negative rate.

Masked Perceptual Loss ($\mathcal{L}_P$): In order to stabilize the training, we added a perceptual loss [5] masked over the clothing regions. This loss penalizes the L_1 distance between the original and inpainted images after being projected into a high dimensional feature space.

Feature Matching Loss ($\mathcal{L}_F$): To further stabilize the training process we penalize high level features on the discriminators [9], by enforcing the generator to match statistics of the original and inpainted images at multiple feature levels of the two discriminators.

Total Loss: The final min-max problem is:

$$G^\star = \arg\min_G \max_D \lambda_I \mathcal{L}_I + \lambda_P \mathcal{L}_P + \lambda_F \mathcal{L}_F \quad (2)$$

where λ_I, λ_P and λ_F are the hyper-parameters that control the relative importance of every loss term and $G^\star$ draws samples from the data distribution.

4 Experimental Evaluation

We next report quantitative and qualitative results for both images and videos. Table 1 provides a quantitative comparison with the state-of-the-art [7, 2, 4] using the Inception Score (IS) [8]. Despite [10] is also a closely related work, its code is not available, preventing its comparison. Our results are consistently better than the other approaches, and very close to the real data IS.

Fig. 4-left depicts results on still images. In some of the examples (e.g. woman with a large coat) there exist large differences between the source and target clothes but the results are still very photo-realistic. Fig. 4-right presents results on image-to-video cloth transfer. In each sequence the left-most column corresponds to two reference cloth images $\mathbf{I}_c$ (source) to be transferred to the target images $\mathbf{X}_t$ displayed on the top row. For every video frame, we show the cloth segmentation estimation $\mathbf{M}_t$ and the output images $\mathbf{Y}_t$ with the transferred clothes. Note that our model

Figure 4: Transferring clothes in images and videos. **Left:** Cloth transfer in still images, between a source $\mathbf{I}_c$ and target $\mathbf{X}_t$. In each case we report the initial estimation $\mathbf{X}'_t$ and the final result $\mathbf{Y}_t$. Missing areas after removing the original cloth and warping the reference cloth are marked in yellow. **Right:** Image-to-video cloth transfer.

Method	mean	std
Pose GAN [7]	2.46	0.80
Pose Variational U-NET [2]	2.79	0.36
VITON [4]	3.11	0.68
Ours $\mathbf{X}'_t$ (only Memory Query)	3.47	0.56
Ours $\mathbf{Y}_t$ (Memory Query + Generator Completion)	**3.94**	**0.89**
Real Data (Upper Bound)	4.21	0.62

Table 1: Quantitative evaluation using the Inception Score [8] metric (the highest the better).

shows remarkable temporally consistent results and robustness to cluttered backgrounds and different body postures. Furthermore, in contrast to previous methods [6, 4], we do not require the person nor the reference cloth to be initialized from a predefined position. This provides our system with a high flexibility towards being applied on unrestricted images from the Internet.

References

[1] R. Alp Gler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 2

[2] P. Esser, E. Sutter, and B. Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018. 3, 4

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 3

[4] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018. 1, 3, 4

[5] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 3

[6] G. Pons-Moll, S. Pujades, S. Hu, and M. Black. Clothcap: seamless 4d clothing capture and retargeting. *TOG*, 2017. 4

[7] A. Pumarola, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer. Unsupervised Person Image Synthesis in Arbitrary Poses. In *CVPR*, 2018. 3, 4

[8] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 3, 4

[9] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 3

[10] M. Zanfir, A.-I. Popa, A. Zanfir, and C. Sminchisescu. Human appearance transfer. In *CVPR*, 2018. 1, 3

[11] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2