Human Synthesis and Scene Compositing

Mihai Zanfir,[3,1] **Elisabeta Oneata,**[3,1] **Alin-Ionut Popa,**[3,1] **Andrei Zanfir,**[1,3] **Cristian Sminchisescu**[1,2]
[1]Google Research [2]Department of Mathematics, Faculty of Engineering, Lund University
[3]Institute of Mathematics of the Romanian Academy
{mihai.zanfir, elisabeta.oneata, alin.popa, andrei.zanfir}@imar.ro, sminchisescu@google.com

Abstract

Generating good quality and geometrically plausible synthetic images of humans with the ability to control appearance, pose and shape parameters, has become increasingly important for a variety of tasks ranging from photo editing, fashion virtual try-on, to special effects and image compression. In this paper, we propose a HUSC (**HU**man **S**ynthesis and Scene **C**ompositing) framework for the realistic synthesis of humans with different appearance, in novel poses and scenes. Central to our formulation is 3d reasoning for both people and scenes, in order to produce realistic collages, by correctly modeling perspective effects and occlusion, by taking into account scene semantics and by adequately handling relative scales. Conceptually our framework consists of three components: (1) a human image synthesis model with controllable pose and appearance, based on a parametric representation, (2) a person insertion procedure that leverages the geometry and semantics of the 3d scene, and (3) an appearance compositing process to create a seamless blending between the colors of the scene and the generated human image, and avoid visual artifacts. The performance of our framework is supported by both qualitative and quantitative results, in particular state-of-the art synthesis scores for the DeepFashion dataset.

Introduction

Generating photorealistic synthetic images of humans, with the ability to control their shape and pose parameters, and the scene background is of great importance for end-user applications such as in photo-editing or fashion virtual try-on, and for a variety of data-hungry human sensing tasks, where accurate ground truth would be very difficult if not impossible to obtain (*e.g.* the 3d pose and shape of a dressed person photographed outdoors).

One way to approach the problem would be to design human models and 3d environments using computer graphics. While the degree of realism increased dramatically in narrow domains like the movie industry, with results that pass the visual Turing test, such synthetic graphics productions require considerable amount of highly qualified manual work, are expensive, and consequently do not scale. In contrast, other approaches avoid the 3d modeling pipeline altogether, aiming to achieve realism by directly manipulating images and by training using large-scale datasets. While this is cheap and attractive, offering the advantage of producing outputs with close to real statistics, they are not nearly as controllable as the 3d graphics ones, and results can be geometrically inconsistent and often unpredictable.

In this work we attempt to combine the relatively accessible methodology in both domains and propose a framework that is able to realistically synthesize a photograph of a person, in any given pose and shape, and blend it veridically with a new scene, while obeying 3d geometry and appearance statistics. An overview is given in fig. 1.

For the first part of human synthesis, given a source image of a person and a different target pose (or more generally, a target 3d body mesh), we want to generate a realistic image of the person synthesized into the new pose. We would like that all the elements included in the person's source layout (either clothing, accessories or body parts) to be preserved or plausibly extended in the synthesis. We propose to learn a dense displacement field, that leverages 3d geometry and semantic segmentation (*e.g.* a blouse *moves* differently than a hat; the leg *moves* differently than the head). This module produces a correction to an initial body displacement field and is trained jointly with the synthesis model within a single end-to-end architecture.

Finally, given an image of a person (real or synthesized) and of a background scene, we want to generate a good quality composite of the two. We argue that a realistic synthesis should consider the physical properties of the scene and of the human body, and also compensate for the different appearance statistics. With that in mind, we propose to blend the foreground image with the background image, at two levels: geometry and appearance. At the geometric level, we want the foreground person, with its associated 3d body model, to respect the 3d space and scale constraints of the scene, be visible according to the scene depth ordering, and be placed on a plausible support surface (*e.g.* floor). At the appearance level, we would like that the two sources blend naturally together, without the undesirable cut-and-paste look. To summarize, our contributions are as follows: *(a)* a realistic human appearance translation task, with state-

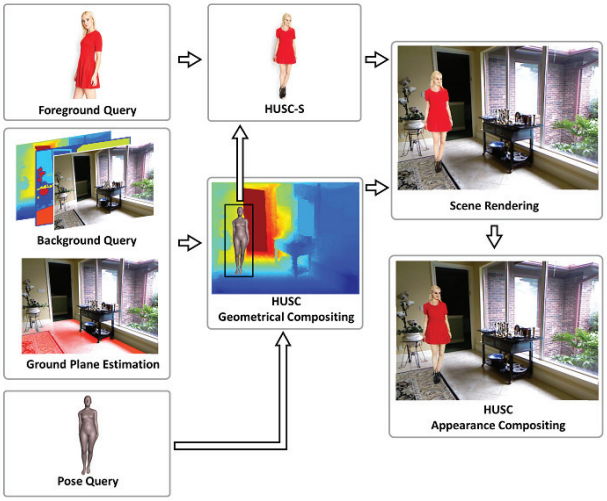

Figure 1: Overview of our full pipeline HUSC. Inputs are a foreground image of a person, a background scene with its associated depth map and semantic labeling, and a target 3d body model. First, we perform ground plane estimation of the background scene. Within the **Geometrical Compositing** stage, we sample a valid 3d location for the target body and perform the associated viewpoint transformation and alignment with the supporting plane normal. The newly updated target body shape, together with the input image encoding the desired appearance, are passed to the human synthesis network, HUSC-S. The resulting synthesized foreground image is rendered in the background scene, by properly accounting for depth ordering constraints. Finally, its appearance is altered by our learned **Appearance Compositing** network in order to produce the final result.

of-the-art results, *(b)* a realistic data augmentation procedure, which allows for the synthesis of complex scenes containing humans, with available pseudo-ground-truth labels such as: pose, shape, segmentation and depth.

Related Work

Human Image Synthesis. An important body of work in the literature is dedicated to image synthesis (Goodfellow et al. 2014; Nguyen et al. 2017; Isola et al. 2017; Yi et al. 2017; Lee et al. 2018; Wang et al. 2018; Luan et al. 2017; Johnson, Alahi, and Fei-Fei 2016), and more specifically to the task of synthesizing photo-realistic images of humans (Zanfir, Popa, and Sminchisescu 2018; Ma et al. 2017; Han et al. 2018; Ma et al. 2018; Siarohin et al. 2018; Lassner, Pons-Moll, and Gehler 2017; Grigorev et al. 2018; Li, Huang, and Loy 2019). Among these, a significant proportion – our work included – have focused on synthesizing humans given a condition image and a desired target pose. In (Esser, Sutter, and Ommer 2018) the authors propose a variational U-Net for conditional image generation. Their method synthesizes images of humans based on 2d pose information and a latent appearance representation learned with a variational auto-encoder. We leverage a richer shape and position representation in the form of a 3d body model and learn a dense correspondence field between the pose and shape of the source person and that of the target. This motion field is extended beyond body regions to clothing and hair. (Dong et al. 2018) learns a transformation grid based on affine and thin-plate spline embedding, which is used to warp the condition image in the desired target position. In contrast, our learned displacement field allows for dense arbitrary transformations. Our work relates to (Siarohin et al. 2018), through the use of deformable skip connections for warping the features maps of the conditioning image, in the location of the desired target pose. However, there are two major differences between our method and the DSCF net proposed in (Siarohin et al. 2018): i) In DSCF net, the image features are warped with an affine transformation obtained through an optimization step both at training and at testing time. ii) The deformable skip connection in DSCF net transform feature-maps, that correspond to coarse 2d body joint activations, while our learned displacement field is densely computed over the entire person layout (body and clothing). Different from (Zanfir, Popa, and Sminchisescu 2018)(HAT), we learn end-to-end the dense correspondence field, coupling it together with the synthesis part of our network. HAT operates in two different stages: one that synthesizes color only on the target body shape, and one that tries to further complete the color on the outside regions. This latter stage is completely detached from the initial source image and is guided only by a target clothing segmentation.

General data augmentation through image synthesis. Combining different sources of synthetic data with real data, and then generating a realistic composition of the both, has been successfully applied to various tasks, such as semi-supervised foreground-background segmentation (Remez, Huang, and Brown 2018; Alhaija et al. 2018; Dwibedi, Misra, and Hebert 2017), object detection (Dvornik, Mairal, and Schmid 2018; Dwibedi, Misra, and Hebert 2017) or 3d object pose estimation (Alhaija et al. 2018). The two cut-and-paste methods (Dwibedi, Misra, and Hebert 2017; Remez, Huang, and Brown 2018) use simple blending techniques, and only for the foreground object, while we propose to learn a blending for both the background and foreground, accordingly. Note that while (Alhaija et al. 2018) and (Dwibedi, Misra, and Hebert 2017) take the 3d geometry of the scene into account, they only consider 3d rigid objects.

Human Synthesis on Backgrounds. Most of the previous works focus solely on the human appearance, and are not concerned with a natural blending within a background image. There are some works that consider synthesizing the entire scene, such as (Varol et al. 2017), where textured human body models are overlaid over random backgrounds. This method does not consider the semantic and the 3d structure of the scene, hence, the human model is often placed in unnatural locations. Moreover, the model's dimensions and illumination do not match the ones of the surrounding environment. Another relevant work is the one of (Balakrishnan et al. 2018), that takes as input an image containing a human and its desired 2d skeleton configuration. The method

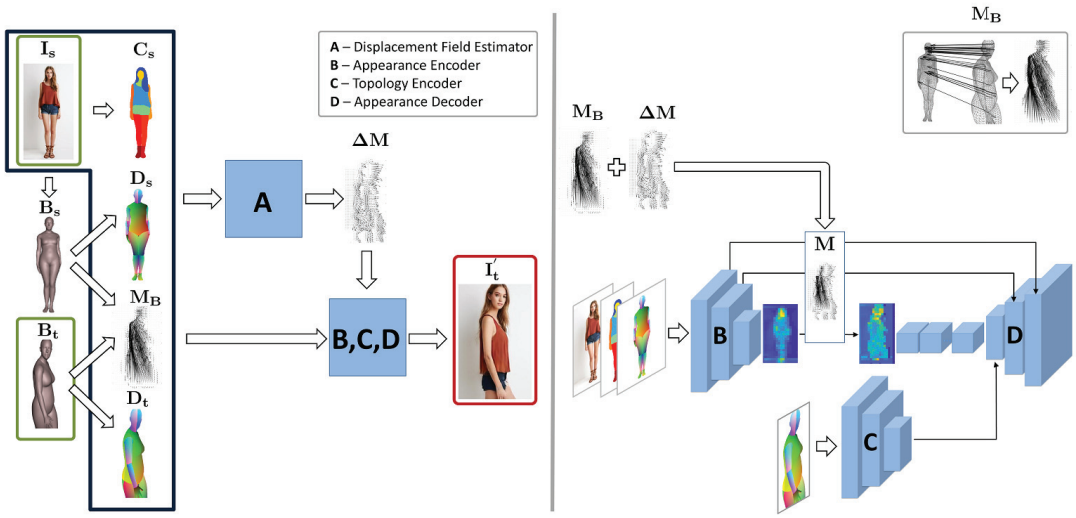

Figure 2: **(Left)** Overview of HUSC-S. Our pipeline receives as input (shown inside green boxes) a source monocular image of a person, I_s, and a desired target 3d body mesh, B_t. We estimate for I_s a 3d body mesh, B_s, and a clothing segmentation, C_s. From B_s and B_t, we can compute a dense body displacement field M_B, and also their respective dense pose representations, D_s and D_t. The **displacement field estimator (A)**, on the top branch, receives as input all the available information, for both source and target. It outputs an update ΔM that is added to M_B to produce the final displacement field, M. **(Right)** Detailed view of HUSC-S. The **appearance encoder (B)** receives as input only the information pertaining to the source. It produces feature maps at different resolutions that are all displaced according to M. We show an example of such a transformation at the lowest resolution. The **topology encoder (C)** operates on the target dense pose, D_t, and is input to the decoder **D** to help guide the synthesis, alongside all the feature maps coming from **B**. The output of our network is the final synthesized image I'_t (shown inside the red box, left).

changes the position of the human to the given pose configuration, while retaining the background scene and inpainting the missing areas.

Human Synthesis

Overview. Given a source RGB image I_s of a clothed person, our goal is to realistically synthesize the same person in a given novel pose, in our case represented by a 3d body model, B_t. Different from previous works that rely mostly on 2d joints to model the pose transformation, one of our key contributions lies in incorporating richer representations, such as 3d body shape and pose which improves the predictions in complicated cases – e.g. determining how the position/occlusion of the body and clothing/hair change with the motion/articulation of the person. Moreover, a 3d body model is required to plausibly place the person in the scene and allows us to have full control over 3d body shape, pose, and the scene location where the synthesized person is inserted.

Our human synthesis model is a conditional - GAN (Wang et al. 2018) that consists of a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$. Given a source image I_s and a condition pose B_t, the task of the generator is to produce an output image $I'_t = \mathcal{G}(I_s)$ of the person in I_s having the body pose B_t. The discriminator's objective is to distinguish between real and generated images while the generator's task is to produce fake images which are able to fool the discriminator.

Starting from an encoder-decoder architecture, our key contribution is the addition of a novel *Displacement Field Estimator* in the generator, that learns to displace features from the encoder, prior to being decoded. Our modified architecture also relies on the semantic segmentation (i.e. clothing and body parts) of the source image, C_s, and on an estimated 3d body model of the source, B_s. An overview of our generator can be seen in fig. 2. We use a multiscale discriminator as in (Wang et al. 2018).

3d Body Model Estimation. For estimating the 3d body model of an image, we use the SMPL model (Loper et al. 2015) and follow a similar procedure to (Zanfir, Marinoiu, and Sminchisescu 2018), but instead apply the method of (Zanfir et al. 2018) for 3d pose estimation. The SMPL model is a parametric 3d human mesh, controlled by pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{24\times 3}$ and shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$, that generate vertices $\mathbf{V}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{6890\times 3}$. We fit this model to predicted 2d keypoints and semantic segmentation, under a full-perspective model with fixed camera intrinsics. We refer to the tuple $(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{T})$ as a model B, where $\mathbf{T}$ is the inferred camera-space translation.

Dense Body Correspondences. Given two body meshes (i.e. a source and a target) we can compute a 3d displacement field M_B^{3d} of the visible surface points on the source $\mathbf{V_s}(\boldsymbol{\theta}_s, \boldsymbol{\beta}_s)$, to the target surface points $\mathbf{V_t}(\boldsymbol{\theta}_t, \boldsymbol{\beta}_t)$ (see fig. 2, top right corner).

$$M_B^{3d} = \mathbf{V_s}(\boldsymbol{\theta}_s, \boldsymbol{\beta}_s) - \mathbf{V_t}(\boldsymbol{\theta}_t, \boldsymbol{\beta}_t) \quad (1)$$

The displacement field is projected in 2d, and encoded onto the target shape, representing the offset from where information is being transferred. We will refer to this final 2d displacement field as $M_B = \mathcal{P}(M_B^{3d})$.

Topological Representation. From a 3d body mesh we obtain a 2d dense pose (Alp Guler, Neverova, and Kokkinos 2018) representation for both the source, D_s, and the target, D_t. The dense pose representation encodes at pixel level, for the visible 3d body points, the semantic body part it corresponds to, alongside the 2d coordinates in the local body part reference system. Each 3d body point will therefore have a unique associated value using this encoding scheme.

Image Synthesis Network. Our image synthesis network generator (HUSC-S) has two novel computational units: a *Displacement Field Estimator*, and a *Topology Encoder*. The former learns an update, ΔM, for the displacement field M_B, which is meant to either correct the warping induced by an erroneous fitting, or capture the motion of structures that fall outside the fitted model (e.g. hair, clothing). The final displacement field is given by $M = \Delta M + M_B$, and is used to place source features at the correct spatial location in the target image, to facilitate decoding and thus, synthesis. This update step is computed on geometric and semantic features, namely: I_s, D_s, D_t, C_s and M_B.

An *Appearance Encoder* module extracts features A_s based on the RGB image I_s, a pre-computed clothing segmentation C_s (using (Gong et al. 2018)) and the source dense pose, D_s. The updated displacement field M is used to warp the appearance features A_s and produce warped image features $A_t^{'}$. These are then passed through a series of residual blocks. The result is concatenated with an encoding of D_t, produced by the *Topology Encoder*, and then passed to an *Appearance Decoder* to generate the synthesized image, $I_t^{'}$. This additional information is useful to the decoding layers, since the underlying body-part topology is exposed to them. We add deformable (i.e. warped by our estimated displacement field) skip connections between the *Appearance Encoder* and the *Appearance Decoder*, to propagate lower level features directly to the decoding phase.

Training. At training time, we are given pairs of images $\{(I_s, I_t)\}$ which are used to learn both the parameters of the discriminator, θ_d, and those of the generator, θ_g, using a combined $L1$ color loss, a perceptual VGG loss (Johnson, Alahi, and Fei-Fei 2016) and a discriminator GAN loss.

$$\mathcal{L} = \mathcal{L}_{GAN} + \lambda \mathcal{L}_{L1} + \gamma \mathcal{L}_{VGG} \tag{2}$$

$$\begin{aligned} \mathcal{L}_{GAN} = \mathbb{E}_{(I_s,I_t)}[\log \mathcal{D}(I_s, I_t)] + \\ \mathbb{E}_{I_s}[\log(1 - \mathcal{D}(I_s, \mathcal{G}(I_s))] \end{aligned} \tag{3}$$

$$\mathcal{L}_{L1} = \sum_{(u,v)} ||I_t(u,v) - I_t'(u,v)||_1 \tag{4}$$

$$\mathcal{L}_{VGG} = \sum_{i=1}^{N} \frac{1}{N_i} \left|\left| VGG^{(i)}(I_t) - VGG^{(i)}(I_t') \right|\right|_1 \tag{5}$$

N represents the number of layers considered from the VGG network, N_i is the dimension of the output features and $VGG^{(i)}(I)$ are the activations of the i^{th} layer on image I.

Human and Scene Compositing

In image compositing, we are interested in combining two visual elements from different sources into a single image, creating the impression that they are actually part of the same scene. In order to create this effect, the addition of the person in the scene should respect the physical laws of the real world, and the appearance statistics of the person (given by illumination, contrast, saturation, resolution) should match those of the scene. To meet our goals, we consider that we have a background image with associated depth and semantic information, and a foreground image of a person (synthesized or real) with an estimated 3d body model (see fig. 1 for an overview).

Geometrical Compositing

We are interested in generating images that respect real world physical constraints, such as: people are (usually) sitting on a ground plane, do not penetrate objects in the scene and have plausible body proportions considering the scale of the environment. In order to estimate a ground plane, we select all 3d points that correspond to the semantic classes associated with a floor plane surface. We fit a plane using a least-squares solution, combined with a RANSAC approach to ensure robustness to outliers. This is necessary to mitigate the noise in the depth signal.

In order to choose a physically plausible position for the body model in the scene, we sample 3d floor locations on a regular grid. To make the model perpendicular to the plane, we compute a rotation matrix that aligns the "up" vector of the model to the plane normal, and multiply it with the original rotation matrix. Also, due the computed camera-space translation, the relative camera orientation will change. Adjusting for both the rotation induced by the plane normal alignment, and the camera point-of-view, involves a rotation of the original SMPL model. Next, we test whether the mesh model collides with other 3d points in the scene. We approximate the body model with simple 3d primitives and check for intersections. If no collisions are detected, we can then place the foreground image. Notice that this is not at the reach of other methods, as they do not have the means to correct their foreground hypotheses based on 3d transformations.

In order to correct the appearance change, associated with these transformations, we run our HUSC-S component for the foreground image and its updated corresponding 3d model. We select the color of a pixel according to the relative depth ordering of 3d points, belonging to both the model and the scene (e.g if the model is behind a desk, some of its body parts will be occluded). An important property of this procedure is its scalability: none of the components requires human intervention, as the whole pipeline is fully automated.

Appearance Compositing

In order to create a natural composition of foreground and background image pairs, we build on a methodology originally proposed in (Tsai et al. 2017) for generating realistic composites of arbitrary object categories over backgrounds. (see fig. 3 for an overview). We learn an adjustment of the color and boundaries of a human composite such that it blends naturally (in a statistical, learned sense) in the scene. For training we use images with humans from COCO. We alter image statistics inside the human silhouette so that it looks unnatural, thus simulating a silhouette pasted on a different background. The network learns to recreate the original image.

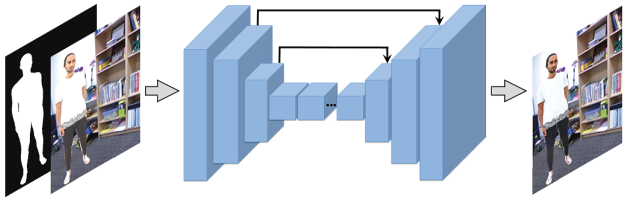

Figure 3: Our Appearance Compositing architecture is represented by an encoder-decoder network inspired by the generator used in (Wang et al. 2018). It takes as input the geometrical composited image and a figure ground segmentation of the person, and corrects his appearance in order to naturally match the scene of insertion.

We make a series of adjustments to the methodology proposed in (Tsai et al. 2017). First, we restrict the problem to generating realistic composites of only humans and backgrounds. We build our model based on a different, more recent network architecture (Wang et al. 2018) and we drop the scene parsing task, since our compositing problem is defined on a specific category (humans). The network's input is a concatenation of the modified image and the binary figure-ground mask (see fig. 3). We add skip connections between the encoder and decoder, such that the network can easily access background features in the decoding stage. We also remove the batch normalization layers, so that the original image statistics are preserved. We train the network by using a combined $L1$ color loss, a perceptual VGG loss (Johnson, Alahi, and Fei-Fei 2016) and a discriminator GAN loss.

Experimental Results

Evaluation metrics. We use several metrics to test the quality of our generated images: the Learned Perceptual Image Patch Similarity metric (LPIPS) (Zhang et al. 2018), the Inception Score (IS) (Salimans et al. 2016) and the Structural Similarity Index (SSIM) (Wang et al. 2004).

Human Synthesis

Datasets. We train our model on the DeepFashion (Inshop Clothes Retrieval Benchmark) (Liu et al. 2016) dataset, which contains $52,712$ in-shop clothes images and around $200,000$ cross-pose/scale pairs. We use the train/test split provided by (Siarohin et al. 2018), containing $101,268$ train and $8,616$ test pairs of the same person in two different poses.

Quantitative Evaluation. In table 1 we provide quantitative results of HUSC-S on the DeepFashion dataset and compare them with previous works. In addition, we perform an ablation study to show the impact of the learned displacement field in the quality of the generated images. Our first baseline (HUSC-S w/o displacement field) is a simplified version that does not use a displacement field and its associated encoder. The source image, I_s, source dense pose D_s, source clothing segmentation C_s, and target dense pose D_t are concatenated and passed directly to the *Appearance Encoder*. As can be seen in table 1, this is a strong baseline, achieving competitive results with previous works under the IS metric, and state-of-the-art results under SSIM. Its strength lies in leveraging more complex pose and shape information in the form of dense pose (previously only used by DPT (Neverova, Alp Guler, and Kokkinos 2018)). In the second baseline (HUSC-S w/ fixed body displacement field), we compute the body displacement field between the 3d body model of the source image and that of the target image, and use it to warp the encoded appearance features of the source. Even if this displacement field is fixed and limited only to body regions, it still improves the results over the first baseline. In the third baseline, we use a dedicated *Displacement Field Estimator* (see fig. 2) to learn a correction of the initial body displacement, that is extended to clothes and hair. The appearance features of the source image are warped with the corrected motion field, right before the residual blocks computation. By doing so, we obtain the largest performance increase, as well as state-of-the-art results on both IS and SSIM scores. This shows the positive impact that the learned displacement field has on the quality of the generated images. The full method (HUSC-S full) adds two elements: i) the deformable skip connections between the encoder and the decoder for faster convergence and ii) the target dense pose in the decoding phase through the *Topology Encoder*, such that information on the position in which the person should be synthesized is explicitly available.

Methods	IS ↑	SSIM ↑
pix2pix(Isola et al. 2017)	3.249	0.692
PG2(Ma et al. 2017)	3.090	0.762
DSCF(Siarohin et al. 2018)	3.351	0.756
UPIS(Pumarola et al. 2018)	2.970	0.747
BodyROI7(Ma et al. 2018)	3.228	0.614
AUNET(Esser, Sutter, and Ommer 2018)	3.087	0.786
DPT(Neverova, Alp Guler, and Kokkinos 2018)	3.61	0.785
SGW-GAN(Dong et al. 2018)	3.446	0.793
DIAF (Li, Huang, and Loy 2019)	3.338	0.778
HUSC-S w/o displacement field	3.3876	0.8035
HUSC-S w/ fixed body displacement field	3.4541	0.8016
HUSC-S w/ learned displacement field	3.6343	0.8049
HUSC-S full	**3.6950**	**0.8135**

Table 1: Evaluation on the DeepFashion dataset. We perform several ablation studies, showing that learning the displacement field, compared to no or fixed body displacement, performs better in terms of both IS and SSIM scores. Our full method obtains state-of-the art results.

Figure 4: Appearance transfer results of a single RGB image into various poses. The first image represents the source, while the others are obtained by synthesizing that person in different poses.

Figure 5: Sample results on the DeepFashion dataset. From left to right: source image, target image, HUSC-S w/ fixed body displacement field, HUSC-S w/ learned displacement field, HUSC-S full. Notice the superior quality of the images synthesized with learned displacement field.

Figure 6: Human synthesis with varying shape parameters. **(Left)** Source image. **(Center)** Synthesized image with same shape parameters as the source. **(Right)** Synthesized image with *larger* shape parameters than the source.

Qualitative Evaluation. A visual comparison of the baselines and the full method can be seen in fig. 5. The first column represents the source image, the second column is the ground truth target image, while the last three columns are outputs of the following methods: HUSC-S w/ fixed body displacement field, HUSC-S w/ learned displacement field and HUSC-S full. Notice the superior quality of the images synthesized with learned displacement (columns 4 and 5) in terms of i) *pose*: the order in which the legs overlap in the first image, ii) *clothing details*: the folds of the dress in the third row look more realistic when using learned motion field, iii) *color/texture preserving*: the color in the first row and the hat in the second row better resemble those of source image and iv) *face details*: in all cases the face of the synthesized person is sharper and better resembles the source person when using the full method. Our method can synthesize high-quality images of persons in a wide range of poses and with a large variety of clothing types, both loose and closely fitted to the body. Moreover, HUSC-S allows altering the underlying 3d pose and shape of the desired target person, while correspondingly adjusting the appearance. Please see fig. 6 for examples of body proportions variations, and fig. 4 for synthesized images of the same person in various poses.

Figure 7: Before and after Appearance Compositing. We learn an adjustment of the color and boundaries of a human composite such that it blends naturally (in a statistical, learned sense) in the scene.

Appearance and Geometrical Compositing

Datasets. For training the Appearance Compositing network, we use the COCO (Lin et al. 2014) dataset due to its large variety of natural images, containing both humans and associated ground truth segmentation masks. For the Geometric Compositing pipeline, we sample various backgrounds from the NYU Depth Dataset V2 (Nathan Silberman and Fergus 2012). This dataset contains $1,449$ RGB images, depicting 464 indoor scenes captured with a Kinect sensor. For each image, the dataset provides corresponding semantic annotations as well as aligned depth images. We consider the semantic classes "floor", "rug", "floor mat" and "yoga mat" in order to infer plausible support surfaces.

Quantitative Evaluation. Table 2 shows the performance of our Appearance Compositing network on images from the COCO validation dataset under the LPIPS metric. We first report the LPIPS metric between the initial images and their randomly perturbed versions (0.0971). Then we show how this error score is considerably improved when we apply our Appearance Compositing network, that uses VGG and L1 loss (from 0.0971 to 0.0588). Note that there is also a considerable reduction in the standard deviation. The re-

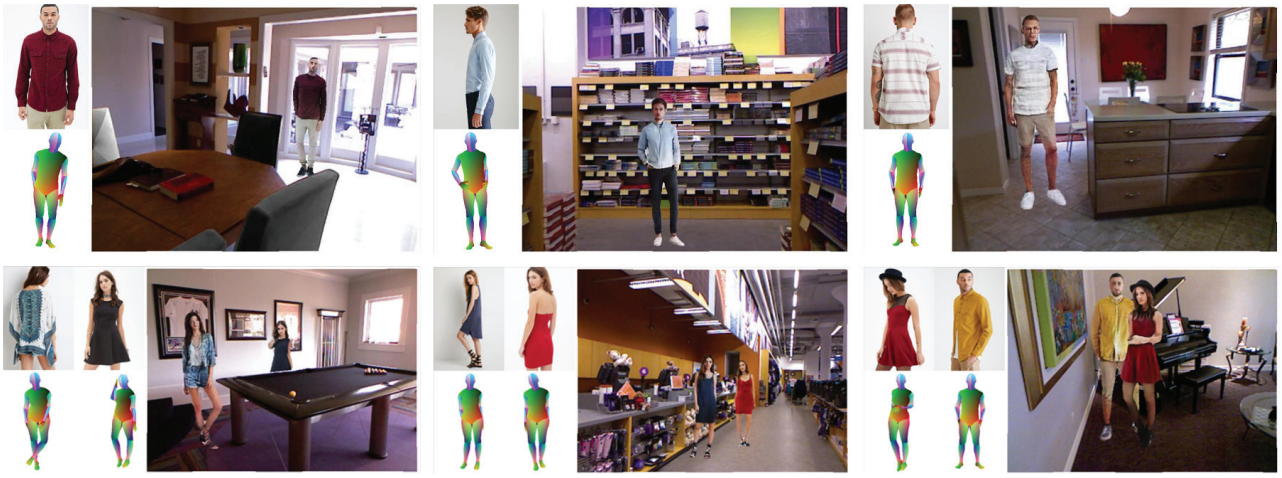

Figure 8: Sample images generated by our framework. For each example, we show the source image, the target 3d body mesh, and a scene with a geometrically plausible placement of the synthesized person. Please note that our framework allows for a positioning behind various objects, and the insertion of multiple people without breaking any geometrical scene properties.

sults further improve if we add a GAN loss to our network (from 0.0588 to 0.0542).

	LPIPS score	
Reference	**Full image** $\downarrow$	**Foreground** $\downarrow$
Perturbed	0.0971 ± 0.0596	0.2569 ± 0.0811
Refined (L1+VGG loss)	0.0588 ± 0.0267	0.1317 ± 0.0514
Refined (L1+VGG+GAN loss)	$\mathbf{0.0542 \pm 0.0243}$	$\mathbf{0.1165 \pm 0.0480}$

Table 2: LPIPS (smaller the better) on 1000 images from Microsoft COCO validation dataset.

In table 3 we report the Inception Score for 1000 composited images of humans on diverse backgrounds, generated by HUSC: with only the Geometrical Compositing, and when the Appearance Compositing module is also added. The score is improved when correcting the appearance of the geometrical composites, which indicate that the images become more realistic. As a comparison baseline, we checked the IS of the background images alone. Note that there is only a less than 10% drop, in the score of the final composite image, as compared to the original (real), background image.

Image set	**Inception Score** $\uparrow$
NYU Depth Dataset V2	6.96
HUSC (Only Geometrical Compositing)	6.23
HUSC	6.33

Table 3: IS for images generated by our method with and without appearance compositing. We also show IS for the original background images selected from the NYU dataset.

Qualitative Examples. In fig. 8 we show examples of images generated by our proposed HUSC framework. For each example, we show the source image, the dense pose associated with the target body model and the resulting synthesized scene. Notice that the persons are naturally blended in the background scene at both geometrical and appearance levels. In fig. 7 we illustrate before and after results for our appearance compositing network. Our method adapts the foreground to the ambient scene context by darkening the appearance, and smooths the boundaries.

Conclusions

We have presented a **HU**man **S**ynthesis and Scene **C**ompositing framework (HUSC) for the realistic and controllable synthesis of humans with different appearance, in novel poses and scenes. By operating in the 3d scene space rather than image space, except for late global image adaptation stages, and by taking into account scene semantics, we are able to realistically place the human impostor on support surfaces, handle scene scales and model occlusion. Moreover, by working with parametric 3d human models and dense geometric correspondences, we can better control and localize the appearance transfer process during synthesis. The model produces pleasing qualitative results and obtains superior quantitative results on the DeepFashion dataset, making it practically applicable for photo editing, fashion virtual try-on, or for realistic data augmentation used for training large scale 3d human sensing models.

Acknowledgments: This work was supported in part by the European Research Council Consolidator grant SEED, CNCS-UEFISCDI (PN-III-P4-ID-PCE-2016-0535, PN-III-P4-ID-PCCF-2016-0180), the EU Horizon 2020 grant DE-ENIGMA (688835), and SSF.

References

Alhaija, H.; Mustikovela, S.; Geiger, A.; and Rother, C. 2018. Geometric image synthesis. In *(ACCV)*.

Alp Guler, R.; Neverova, N.; and Kokkinos, I. 2018. Densepose: Dense human pose estimation in the wild. In *(CVPR)*.

Balakrishnan, G.; Zhao, A.; Dalca, A. V.; Durand, F.; and Guttag, J. 2018. Synthesizing images of humans in unseen poses. In *(CVPR)*.

Dong, H.; Liang, X.; Gong, K.; Lai, H.; Zhu, J.; and Yin, J. 2018. Soft-gated warping-gan for pose-guided person image synthesis. In *(NeurIPS)*, 474–484.

Dvornik, N.; Mairal, J.; and Schmid, C. 2018. Modeling visual context is key to augmenting object detection datasets. In *(ECCV)*, 364–380.

Dwibedi, D.; Misra, I.; and Hebert, M. 2017. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *(CVPR)*, 1301–1310.

Esser, P.; Sutter, E.; and Ommer, B. 2018. A variational u-net for conditional appearance and shape generation. In *(CVPR)*.

Gong, K.; Liang, X.; Li, Y.; Chen, Y.; Yang, M.; and Lin, L. 2018. Instance-level human parsing via part grouping network. In *(ECCV)*.

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

Grigorev, A.; Sevastopolsky, A.; Vakhitov, A. T.; and Lempitsky, V. S. 2018. Coordinate-based texture inpainting for pose-guided image generation. *CoRR* abs/1811.11459.

Han, X.; Wu, Z.; Wu, Z.; Yu, R.; and Davis, L. S. 2018. Viton: An image-based virtual try-on network. In *(CVPR)*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *(CVPR)*.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *(ECCV)*.

Lassner, C.; Pons-Moll, G.; and Gehler, P. V. 2017. A generative model of people in clothing. In *(ICCV)*, 853–862.

Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; Singh, M. K.; and Yang, M.-H. (2018). Diverse image-to-image translation via disentangled representations. In *(ECCV)*.

Li, Y.; Huang, C.; and Loy, C. C. 2019. Dense intrinsic appearance flow for human pose transfer. In *(CVPR)*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *(ECCV)*.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *(CVPR)*.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* 34(6):248.

Luan, F.; Paris, S.; Shechtman, E.; and Bala, K. 2017. Deep photo style transfer. In *(CVPR)*, 4990–4998.

Ma, L.; Jia, X.; Sun, Q.; Schiele, B.; Tuytelaars, T.; and Van Gool, L. 2017. Pose guided person image generation. In *(NIPS)*, 406–416.

Ma, L.; Sun, Q.; Georgoulis, S.; Van Gool, L.; Schiele, B.; and Fritz, M. 2018. Disentangled person image generation. In *(CVPR)*, 99–108.

Nathan Silberman, Derek Hoiem, P. K., and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *(ECCV)*.

Neverova, N.; Alp Guler, R.; and Kokkinos, I. 2018. Dense pose transfer. In *(ECCV)*, 123–138.

Nguyen, A.; Clune, J.; Bengio, Y.; Dosovitskiy, A.; and Yosinski, J. 2017. Plug & play generative networks: Conditional iterative generation of images in latent space. In *(CVPR)*, 4467–4477.

Pumarola, A.; Agudo, A.; Sanfeliu, A.; and Moreno-Noguer, F. 2018. Unsupervised person image synthesis in arbitrary poses. *(CVPR)*.

Remez, T.; Huang, J.; and Brown, M. 2018. Learning to segment via cut-and-paste. In *(ECCV)*, 37–52.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *(NIPS)*, 2234–2242.

Siarohin, A.; Sangineto, E.; Lathuilière, S.; and Sebe, N. 2018. Deformable gans for pose-based human image generation. In *(CVPR)*.

Tsai, Y.-H.; Shen, X.; Lin, Z.; Sunkavalli, K.; Lu, X.; and Yang, M.-H. 2017. Deep image harmonization. In *(CVPR)*.

Varol, G.; Romero, J.; Martin, X.; Mahmood, N.; Black, M. J.; Laptev, I.; and Schmid, C. 2017. Learning from synthetic humans. In *(CVPR)*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: From error measurement to structural similarity. *(TIP)*.

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

Yi, Z.; Zhang, H. R.; Tan, P.; and Gong, M. 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In *(ICCV)*, 2868–2876.

Zanfir, A.; Marinoiu, E.; Zanfir, M.; Popa, A.-I.; and Sminchisescu, C. 2018. Deep network for the integrated 3d sensing of multiple people in natural images. In *(NIPS)*.

Zanfir, A.; Marinoiu, E.; and Sminchisescu, C. 2018. Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes – The Importance of Multiple Scene Constraints. In *(CVPR)*.

Zanfir, M.; Popa, A. I.; and Sminchisescu, C. 2018. Human appearance transfer. In *(CVPR)*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *(CVPR)*.