Understanding Pose and Appearance Disentanglement in 3D Human Pose Estimation

Krishna Kanth Nakka[1][0000−0002−2381−6593] and
Mathieu Salzmann[1,2][0000−0002−8347−8637]

[1] CVLab, EPFL, Switzerland
[2] ClearSpace, Switzerland
{krishna.nakka, mathieu.salzmann}@epfl.ch

Abstract. As 3D human pose estimation can now be achieved with very high accuracy in the supervised learning scenario, tackling the case where 3D pose annotations are not available has received increasing attention. In particular, several methods have proposed to learn image representations in a self-supervised fashion so as to disentangle the appearance information from the pose one. The methods then only need a small amount of supervised data to train a pose regressor using the pose-related latent vector as input, as it should be free of appearance information. In this paper, we carry out in-depth analysis to understand to what degree the state-of-the-art disentangled representation learning methods truly separate the appearance information from the pose one. First, we study disentanglement from the perspective of the self-supervised network, via diverse image synthesis experiments. Second, we investigate disentanglement with respect to the 3D pose regressor following an adversarial attack perspective. Specifically, we design an adversarial strategy focusing on generating natural appearance changes of the subject, and against which we could expect a disentangled network to be robust. Altogether, our analyses show that disentanglement in the three state-of-the-art disentangled representation learning frameworks if far from complete, and that their pose codes contain significant appearance information. We believe that our approach provides a valuable testbed to evaluate the degree of disentanglement of pose from appearance in self-supervised 3D human pose estimation.

1 Introduction

Monocular 3D human pose estimation has been at the heart of computer vision research for decades, and tremendous results can now be achieved in the supervised learning setting [22, 14, 15, 27, 38, 29, 23, 37, 28, 33, 21]. Unfortunately, obtaining 3D pose annotations for real images remains very expensive, particularly in the wild. As such, self-supervised learning approaches have received an increasing attention in the past few years [32, 31, 12, 5]. One of the common factors across all these methods is their aim to learn a latent representation of the image that disentangles the person's pose from their appearance. In practice, as shown in Figure 1, this has been achieved by leveraging access to either multiple views [31, 32] or video sequences [5, 12] during training. In either case,

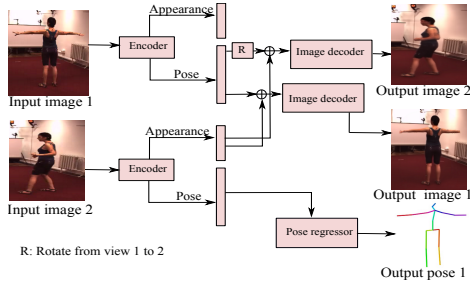


Fig. 1. Disentanglement-based Representation Learning. Given a reference frame and another frame from either a different view or a different time instant, an encoder learns a representation separated into two components, appearance and pose, in a self-supervised fashion. A pose regressor is then trained using limited annotated data to map the latent pose vector to a 3D human pose.

one then only needs access to a small amount of supervised data to effectively train a pose regressor from the pose-related portion of the latent code to the actual 3D pose, because this portion of the latent code should in theory contain only pose-relevant information.

Despite the impressive progress of these self-supervised 3D human pose estimation methods, several fundamental questions about their learnt representations remain unanswered. For example, to what extent are the pose and appearance latent vectors disentangled? Do these two representations contain truly complementary information, or do they share some signal? How do the different sources of self-supervision, i.e., multiple views or temporal information, affect the disentanglement of these representations?

In this paper, we seek to provide a deeper understanding of such disentangled representations by analyzing the resulting latent spaces in two ways. First, we study the disentanglement of the latent pose and appearance vectors with respect to the self-supervised representation learning network. In this context, we analyze both the images synthesized by altering the appearance codes in different ways, and the influence on pose and appearance of different channels in the latent pose codes. Second, we investigate the disentanglement with respect to the supervised 3D pose regressor. To this end, we follow an adversarial attack strategy, aiming to modify the subject's appearance so as to affect the regressed 3D pose. However, instead of exploiting a standard adversarial attack technique [20, 18, 10], against which disentangled pose networks were never meant to be robust, we design a dedicated framework that should be much more favorable to such networks. Specifically, we seek to alter only the latent appearance vector so as to affect the 3D pose regressed from the latent pose vector extracted from the image synthesized using the modified appearance vector with the original pose one.

Our experiments on the state-of-the-art disentangled representation learning frameworks, NSD [31], CSSL [12] and DRNet [5], evidence that, across the board, *disentanglement is not complete and the pose codes of these frameworks contain appearance information.* Our work provides the tool to study the effectiveness of different disentanglement-based training strategies and will serve as a valuable testbed to analyze the extent of disentanglement in future frameworks.

Contributions. To summarize, our contributions are twofold. (1) We systematically analyze the latent pose and appearance representations in several representative disentangled networks. Our experiments lead to an interesting find-

ing that the latent pose vectors contain almost all of the subject's appearance information. (2) We introduce an adversarial strategy to understand the disentanglement of 3D pose from natural appearance changes. Our code and trained networks will be made publicly available upon acceptance.

2 Related Work

Disentanglement-based 3D Human Pose Estimation. Disentangling pose and appearance in 3D pose estimation was first proposed in DRNet [5], where a discriminator was employed to distinguish if the time-varying features from two images represented the same subject or not. Furthermore, the distance between the time-invariant, i.e., appearance, component of one subject at two different time instants was minimized, and the time-varying pose features were encouraged to be indistinguishable across subjects, thereby ensuring that appearance information did not leaked into the pose features. In [31, 32], disentanglement was achieved via the use of multiple views during training, leveraging the intuition that, for one subject, the pose features extracted from one view and rotated to a different view at the same time instant should be the same as those directly extracted from that view, and the appearance features at different time instants should be similar so as not to contain pose information. More recently, [12] designed a contrastive loss to force the latent appearance features in temporally-distant frames to remain close while encouraging the pose features to be different from each other. All these methods learn the disentangled representation from unsupervised data, and then train a shallow regressor to predict 3D pose from the pose latent vector using a limited amount of pose labels. In this work, we study how disentangled the appearance and pose features extracted by these methods truly are. To this end, we provide analyses based on diverse image synthesis experiments and on adversarial attacks.

Adversarial Attacks. Deep neural networks were first shown to be vulnerable to adversarial examples in [36]. Following this, several attacks have been proposed, using either gradient-based approaches [10, 18] or optimization-based techniques [3, 26, 25, 4, 7]. To study the disentanglement of pose and appearance in 3D human pose estimation, we seek to analyze if appearance changes can affect the regressed 3D pose. In principle, we could use any of the above-mentioned attack strategy to do this. However, they offer no control on the generated perturbations, and thus could potentially incorporate structures that truly suggest a different pose. In other words, the disentangled networks cannot be expected to be robust to such attacks. Therefore, we design an attack strategy to which they can be expected to be robust. Specifically, we synthesize an image by modifying only the appearance code of the network of interest, and show that the 3D pose regressed from that image will typically differ from the original one. Our attacks can be thought of as inconspicuous ones, as the generated image looks natural, with only appearance changes to the subject. Other works [41, 16, 30, 35, 2, 34] have designed strategies to generate realistic adversarial images, typically focusing on face recognition datasets and using GANs [9, 24, 1]. Our approach nonetheless fundamentally differs from those in both methodology and context;

our main goal is not to attack disentangled 3D human pose networks but to study their level of disentanglement. Therefore, we design an attack strategy that is most favorable for these networks, and against which they can be expected to be naturally robust.

Measuring Disentanglement. In other contexts than human pose estimation, several works have proposed metrics to quantify the degree of disentanglement of latent vectors [8, 6, 19]. These methods are of course also applicable to the self-supervised learning frameworks that we will analyze, and we will report these metrics in our experiments. However, these metrics do not provide any understanding of where disentanglement fails. This is what we achieve with our diverse analyses.

3 Disentangled Human Pose Estimation Networks

Given an image as input, 3D human pose estimation aims to predict the 3D positions of J body joints, such as the wrists, elbows, and shoulders. When no annotations are available for the training images, an increasingly popular approach consists of learning a latent representation that disentangles appearance from pose in a self-supervised fashion. Here, we review disentanglement-based 3D human pose estimation frameworks that we will analyze in Sections 4 and 5.

Existing disentanglement-based frameworks essentially all follow the same initial steps. The input image $\mathbf{I}$ is first passed through a spatial transformer network $\mathcal{S}$ to extract the bounding box corresponding to the human subject. An encoder E then takes the cropped bounding box $\mathbf{I}_c$ as input and outputs a latent vector $\mathbf{h}$ comprising two components, that is $E : \mathbf{I}_c \rightarrow [\mathbf{h}_a, \mathbf{h}_p]$. The first component, $\mathbf{h}_a$, aims to encode the subject's appearance while the second, $\mathbf{h}_p$, should represent the subject's pose. The networks are trained without any 3D pose annotations, and thus supervision is achieved via image reconstruction. Specifically, a decoder D takes the complete the latent vector $\mathbf{h}$ as input and and outputs a reconstructed version of the cropped image $\tilde{\mathbf{I}}_c$, with an additional mask $\mathbf{M}$ corresponding to the subject's silhouette. The cropped image is further merged with a pre-computed background image $\mathbf{B}$ to obtain the final reconstructed input image $\tilde{\mathbf{I}}$.

The main difference between existing frameworks lies in the way they encourage the disentanglement of pose and appearance. Specifically, the different frameworks train the encoder E and decoder D as follows:

NSD [31]. The neural scene decomposition (NSD) approach leverages the availability of multiple views during training. Given a pair of images from two views at time t, NSD passes one image to the encoder to obtain an appearance vector $\mathbf{h}_a^t$ and a pose vector $\mathbf{h}_p^t$. The pose vector $\mathbf{h}_p^t$, shaped as a 3D point cloud, is rotated to the second view using the ground-truth camera calibration between the two views to obtain a transformed pose vector $\mathbf{h}_{p,r}^t$. Furthermore, to factor out appearance from pose, NSD replaces the appearance vector $\mathbf{h}_a^t$ by an appearance vector $\mathbf{h}_a^{t_1}$ of the same subject at a different time instant t_1. The decoder D then takes as input $\mathbf{h} = [\mathbf{h}_{p,r}, \mathbf{h}_a^{t_1}]$ and aims to reconstruct the image from the second view at time t.

CSSL [12]. Instead of using multiple views, contrastive self-supervised learning (CSSL) exploits temporal information from videos to learn a latent representation of pose and appearance. To achieve disentanglement, CSSL encourages the distance between the latent pose vectors $\mathbf{h}_p^{t_1}$ and $\mathbf{h}_p^{t_2}$ of two frames, t_1 and t_2, to reflect their temporal distance. Furthermore, similarly to NSD, CSSL swaps the appearance vectors $\mathbf{h}_a^{t_1}$ and $\mathbf{h}_a^{t_2}$ of the two video frames when performing image reconstruction so as to force them to learn time-invariant information, thus encoding appearance.

DRNet [5]. The disentangled representation network (DRNet) uses a similar strategy to that of CSSL, consisting of randomly choosing two temporal frames, t_1 and t_2, from a video. However, DRNet aims to achieve disentanglement in two ways: (1) By minimizing the distance between the two appearance vectors $\mathbf{h}_a^{t_1}$ and $\mathbf{h}_a^{t_2}$; and (2) by exploiting an adversarial network to make the pose vector $\mathbf{h}_p$ independent of the subject's appearance. Specifically, this is achieved by training the additional discriminator to output the subject's identity given the pose vector as input, and training the encoder E in an adversarial fashion to fool the discriminator.

Once trained on a large corpus of unannotated images in a self-supervised manner, the frameworks discussed above employ a 2 layer pose regressor $\phi : \mathbf{h}_p \rightarrow \mathbf{q}$ to predict the 3D pose $\mathbf{q}$ from the latent pose vector $\mathbf{h}_p$. This pose regressor is trained with a small amount of supervised data, while freezing the weights of the encoder. Due to space limitations, we provide additional details about training in the supplementary material.

4 Disentanglement w.r.t. the Self-Supervised Network

In this section, we study the disentanglement of pose and appearance within the self-supervised representation learning network itself. To this end, we first analyze the impact of the latent appearance vector on the images synthesized by the network's decoder. We then turn to investigating the influence on pose and appearance of different channels in the latent pose vector.

4.1 Effect of the Appearance Vector on Synthesized Images

Our first analysis consists of visualizing the images generated by the network's decoder. In particular, we leverage the intuition that, if the pose and appearance vectors were disentangled, altering the appearance vector while keeping the pose one fixed should yield images with a different subject's appearance but the same pose. We investigate this via the two strategies discussed below.

First, we synthesize novel images by mixing the appearance and pose information from two subjects, S8 and S7. The top two rows of Figure 2 show the images synthesized with DRNet[3] *without mixing the appearance vectors; these images look similar to the original ones, depicting two clearly different subject's appearances.* By contrast, the images in the third to fifth row of the figure, obtained

[3] Similar images for the other networks are provided in the supplementary material.

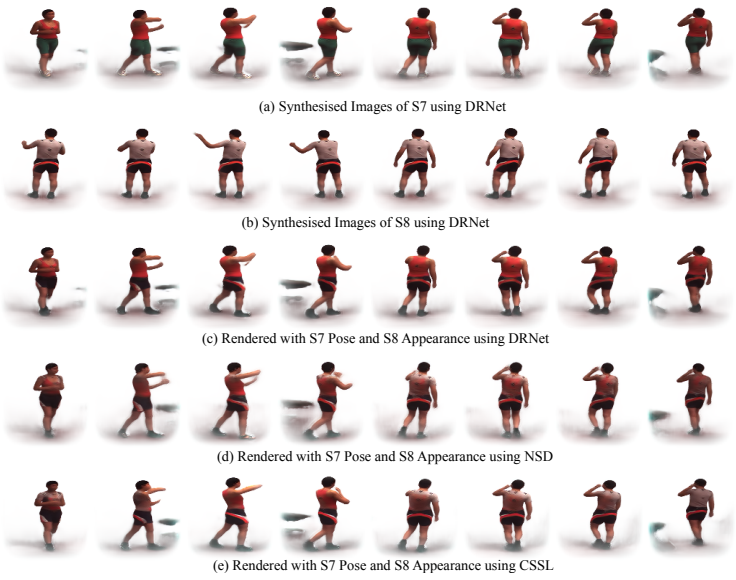


Fig. 2. Synthesizing novel images. We take the pose information from S7 (first row) and the appearance information from S8 (second row) and synthesize novel images in the third, fourth and fifth rows using DRNet, CSSL and NSD. The synthesized images retain some appearance information (red shirt) of S7 although we only use S7's pose code in the synthesis.

by using S7's pose vector and S8's appearance one, still contain appearance information of S7. This is particularly the case for the images synthesized using DRNet and NSD, in which the subject's shirt has taken the red color of that of S7, although we use only S7's pose code in the synthesis process. CSSL is less subject to such failures, but they nonetheless occur in some cases, such as in the third and fourth columns.

As a second experiment, we replace the appearance vector with a zero vector. We then combine this zero appearance vector with the pose vector obtained from the original image shown in the first row of Figure 3. As can be seen from the second row, even though we use the same zero appearance vector to generate images of different subjects, the synthesized images retain almost all the appearance information of the original images, except near the head region.

Both of these experiments evidence that the pose code contains a significant amount of appearance information and that the disentanglement is thus not complete. Nevertheless, both experiments also show that modifying the appearance code indeed does not impact the subject's pose in the synthesized image. To further verify whether the appearance codes are truly free of pose information, we visualize the appearance codes of all images of a S7 using t-SNE in Figure 4. The resulting plot shows nicely-separated clusters, which can be observed to correspond to action categories. This suggests that, although modifying the appearance code does not visually change the subject's pose in the synthesized

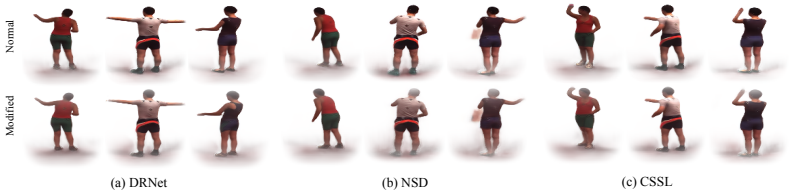


Fig. 3. Replacing the appearance code with a fixed zero vector. In the first row, we show the original synthesized images for three subjects on different networks. In the second row, we set the values in the appearance vector to zero and use the same pose vectors as in the first row. Despite using the same zero appearance vector for all subjects, the outputs do not appear similar in content and instead retain almost all the appearance information of the original images.

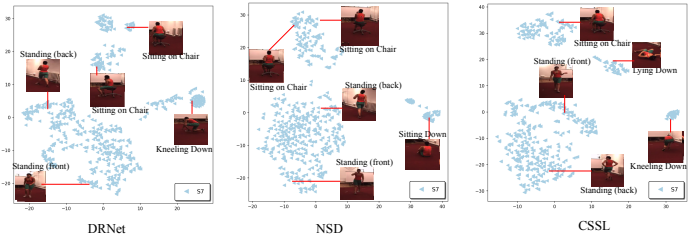


Fig. 4. tSNE visualization of appearance codes. The appearance codes of images from same subject S7 are clustered according to the action performed by the subject. This indicates that the appearance code still contains information about the pose. Best viewed in color and zoomed in.

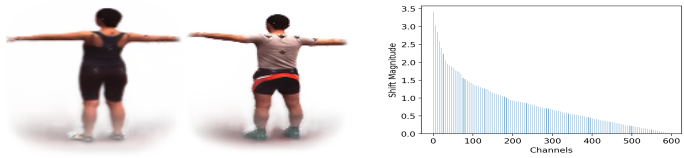
images, the appearance codes still contain information about the subject activity, and thus about their pose.

4.2 Effect of the Pose Vector on Synthesized Images

In this section, we study the impact of the pose vector on the synthesized images and further provide evidence of the presence of appearance information in the pose code. To this end, we identify channels encoding appearance information in the pose code. Our approach is based on the idea that two images depicting different subjects in similar poses should ideally have similar latent pose codes. The channels that have large differences therefore indicate the presence of appearance information.

To illustrate this, we use the two images shown in Figure 5(a) and plot the absolute difference between the corresponding pose codes obtained by NSD[4] in Figure 5(b), ordering the channels by the magnitude of the difference. The latent pose indeed disagree in many channels. We define the probability of a channel to encode appearance information to be proportional to the absolute pose vector difference for that channel. Below, we then analyze the effect of altering the K channels with highest or lowest appearance probability.

[4] Similar plots for the other networks are provided in the supplementary material.



(a) Two images with similar poses but different appearances.

(b) Absolute difference between the pose codes sorted by magnitude.

Fig. 5. Detecting appearance channels in the pose latent vector. We take images depicting different subjects in a similar pose, for which we could expect the pose codes to be close. However, as shown on the right, the latent pose vectors obtained by NSD contain channels with large differences, likely to encode appearance information.

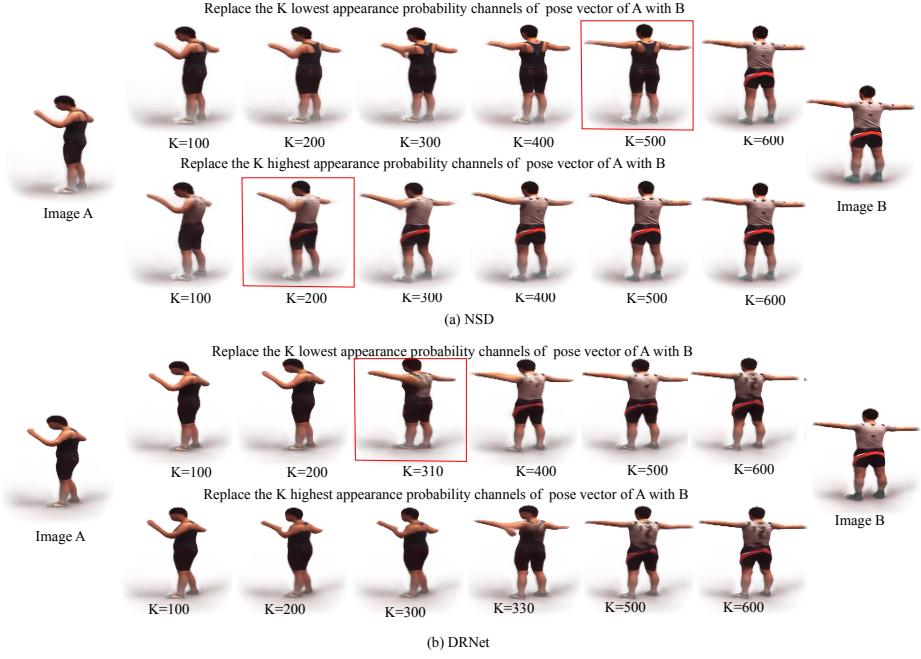


Fig. 6. Influence of the pose code channels. To synthesize the images in the middle portion of the figure, we take the appearance code corresponding to image A, and vary the pose code in two ways. Specifically, in the top (or bottom) portion of the figure, we replace the K channels with lowest (or highest) appearance probability with the corresponding ones from the pose code extracted from image B. (a) For NSD, replacing the $K = 500$ lowest appearance probability channels yields an image (highlighted with a red box) depicting B's pose and A's appearance. Similarly, replacing the $K = 200$ highest appearance probability channels produces B's appearance and A's pose. (b) We observe similar trends for DRNet, although the separation of appearance and pose inside the pose code is not as clear as for NSD.

To this end, we take two images A and B, as shown in the left and right ends of Figure 6, fix the appearance code as that of A. We then replace the channels with either the K lowest or highest appearance probability in the pose

code of A with the corresponding values from the pose code of B. Note that all disentangled networks have a pose code of dimension 600, and therefore $K = 600$ means replacing all the channels of the pose vector.

As shown in Figure 6(a) for NSD, by replacing the $K = 500$ lowest appearance probability channels yields an image (highlighted with a red box) with A's appearance and B's pose. Furthermore, replacing the $K = 200$ highest appearance probability channels synthesizes an image with B's appearance and A's pose. Both these results indicate that the top 100-200 highest probability appearance channels in the pose code indeed encode the appearance information for NSD. It is worth noting that with $K = 600$ the image depicts both the pose and appearance of B, confirming our previous experiments in Figure 2.

Figure 6(b) for DRNet shows the channels are not as clearly separated in pose and appearance ones in this method. Nevertheless, the pose codes still combines pose and appearance information. We present similar analysis and visualizations for CSSL in the supplementary material.

5 Disentanglement w.r.t. the 3D Pose Regressor

The previous set of analyses have focused on the self-supervised representation learning networks themselves, evidencing that the latent pose vector is contaminated with appearance information. Here, we further investigate the disentanglement w.r.t. the supervised 3D human pose regressor, which takes the latent pose vector as input. Note that, since the 3D pose regressor is disassociated from the appearance vector at network level, studying the appearance and pose vector disentanglement in this context is not straightforward. Therefore, we consider the pose estimation network comprised of the self-supervised encoder and the supervised decoder as a standalone network and study the effects of the input image appearance on its 3D pose output. To this end, we introduce an adversarial perturbation strategy that explicitly focuses on modifying only the appearance information in the input image. Below, we first describe our attack framework, and then analyze its effects on the disentangled pose estimation networks.

5.1 Appearance-only Attack Framework

Our goal is to perturb only the subject's appearance in the input image; perturbing the image such that the subject's pose visually changes would of course make the pose regressor output a different pose but would not allow us to verify the disentanglement of pose and appearance. To enforce such a constraint on our perturbations, we follow a strategy that, intuitively, should constitute a weak attack and thus be favorable to the disentangled network. Specifically, we only perturb the latent appearance vector, which we combine with the *original* pose one to generate an adversarial image. We then extract a new latent pose vector from this image and predict the 3D human pose from it. If the pose regressor could discard the appearance information, it would thus not be affected by this perturbation.

As shown in Algorithm 1, we generate an adversarial image $\mathbf{I}_{adv}$ as using a generator network G. In practice, we take G to be either the disentangled network of interest or another disentangled network, and we will report results with both strategies. First, we pass the original input image to the generator's spatial transformer G_s and extract the cropped image $\mathbf{I}_c$ using the resulting bounding box. We then encode the cropped image $\mathbf{I}_c$ into an initial latent pose vector $\tilde{\mathbf{h}}_p^0$ and latent appearance vector $\tilde{\mathbf{h}}_a^0$ using the generator's encoder G_e. The combined latent vector $\tilde{\mathbf{h}} = [\tilde{\mathbf{h}}_a^0, \tilde{\mathbf{h}}_p^0]$ is then passed as input to the generator's decoder G_d, which outputs the reconstructed image $\tilde{\mathbf{I}}_c^0$ and a mask $\mathbf{M}^0$. The cropped output $\tilde{\mathbf{I}}_c^0$ is then combined with the pre-computed background image $\mathbf{B}$ to resynthesize an image $\mathbf{I}_{adv}^0$ at full resolution. This image then acts as input to the target pose estimation network, which encompasses an encoder E, that may differ from the generator one G_e, and a pose regressor. This forward pass produces an initial pose estimate $\phi(\mathbf{h}_p^0)$. Note that the output of the target network given $\mathbf{I}_{adv}^0$ as input has empirically a small mean per-joint position error (MPJPE) of around 20 mm with respect to the prediction $\mathbf{q}$ obtained from the original image $\mathbf{I}$. This is because, at this point, no attack has been performed.

To attack only the subject's appearance in the adversarial input, we fix the pose vector $\tilde{\mathbf{h}}_p = \tilde{\mathbf{h}}_p^0$ to generate images of depicting the subject in their original pose. Furthermore, we also fix the mask to its initial value $\mathbf{M} = \mathbf{M}^0$. We then compute an appearance-only perturbation by optimizing the latent appearance vector $\tilde{\mathbf{h}}_a$ in an iterative manner until it either achieves an MPJPE error with respect to the original prediction $\mathbf{q}^0$ higher than a threshold, or reaches a maximum number of iterations. Note that our previous set of experiments in Section 4 have evidenced that modifying the appearance vector does not change the observed subject's pose, which validates our use of the network's decoder to generate the appearance-modified image.

5.2 Appearance-only Attack Results

Qualitative Results. In Figure 7, we visualize the results of different models on the attacked images. For all disentangled representation frameworks, small changes in appearance produce wrong predictions. In particular, as shown in the third row, a small change in the shirt color leads to a completely different pose for all models. This demonstrates that the pose estimation network is dependent on the subject's appearance in the input image that its intermediate latent pose vector is not completely disentangled from appearance.

Quantitative Study. We provide the results of our appearance-only attacks in Table 1 using the network decoder as the generator. We report the MPJPE at the initial iteration and after the attack for each subject. Specifically, the initial error corresponds to the error between the predictions obtained from the original image $\mathbf{I}$ and from the synthesized image $\mathbf{I}_{adv}^0$, without any latent attack. It is around 21.8 mm on average. This shows that the generator faithfully reconstructs the input image and can therefore be employed to perform the attack. After the attack, the performance decrease across all the disentangled models In other

Algorithm 1 Appearance-only attacks

Require: $\mathbf{I}$: Input image, G: Pre-trained generator (with spatial transformer G_s, encoder G_e and decoder G_d), S: Target spatial transformer, E: Target encoder, D: Target image decoder, ϕ: Target pose regressor

$\mathbf{I}_c \leftarrow G_t(\mathbf{I})$, $[\tilde{\mathbf{h}}_a^0, \tilde{\mathbf{h}}_p^0] \leftarrow G_e(\mathbf{I}_c)$
$\mathbf{I}_{adv}^0 = G_d(\tilde{\mathbf{h}}_a^0, \tilde{\mathbf{h}}_p^0), [\mathbf{h}_a^0, \mathbf{h}_p^0] \leftarrow E(S(\mathbf{I}_{adv}^0))$
$[\mathbf{h}_a, \mathbf{h}_p] \leftarrow E(S(\mathbf{I}))$,
$\mathbf{q} \leftarrow \phi(\mathbf{h}_p)), \text{error}_0 = \left\|\mathbf{q} - \phi(\mathbf{h}_p^0)\right\|^2$
$i \leftarrow 1$
while do $\text{error}_i \leq$ min. error and $i \leq$ max. iterations
    $\mathbf{I}_{adv}^i \leftarrow G_d(\tilde{\mathbf{h}}_a^i, \tilde{\mathbf{h}}_p^0)$
    $[\mathbf{h}_a^i, \mathbf{h}_p^i] \leftarrow E(S(\mathbf{I}_{adv}^i))$
    $\text{error}_i \leftarrow \left\|\mathbf{q} - \phi(\mathbf{h}_p^i)\right\|^2$
    $\tilde{\mathbf{h}}_a^{i+1} \leftarrow \text{BackProp}\,\{\text{error}_i\}$
    $i \leftarrow i + 1$
end while
return $\mathbf{I}_{adv} = \mathbf{I}_{adv}^i$

Fig. 7. Apperance-only Attack Examples. Given an input image **(a)** with ground-truth pose **(d)**, we first reconstruct **(b)** the images using a generator. By optimizing the latent appearance vector, we obtain an adversarial image **(c)** that aims to fool the pose regressor so that it outputs a 3D pose **(f)** that differs significantly from the original predictions **(e)**.

words, all models are vulnerable to our appearance-based attacks and typically reach an MPJPE of at least 175 mm. This indicates that the latent pose vector $\mathbf{h}_p$ is not invariant to appearance changes and therefore that the appearance-pose disentanglement is not complete. We provide ablative study using the same NSD decoder as the generator for all disentangled networks in the supplementary material.

To further evaluate quantitatively the sensitivity of a disentangled network to our appearance-only attacks, we computed three image-based metrics, Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Mean Square Error (MSE), to compare the attacked images with those synthesized with the original framework. As shown in Table 2, the three metrics indicate that the images obtained by attacking DRNet are more similar to the original synthesized ones than those obtained by attacking NSD or CSSL. This suggests that DRNet can be attacked with smaller changes, and thus contains more appearance information in its pose vectors.

Altogether, our experiments evidence that disentangling pose and appearance in an unsupervised manner for 3D human pose estimation remains far from being solved. Our attacks thus provide a valuable testbed to valuate the effectiveness of future disentanglement-based frameworks.

Subject	NSD		DRNet		CSSL		Average	
	Initial	Final	Initial	Final	Initial	Final	Initial	Final
S1	21.0	179.7	23.9	169.7	21.5	176.9	21.6	174.2
S5	19.6	180.0	14.1	166.7	25.3	186.5	19.6	177.1
S6	22.3	179.8	23.5	177.9	26.8	196.7	23.4	184.7
S7	18.8	179.2	17.6	177.5	24.1	191.8	20.3	182.3
S8	16.8	178.6	21.7	198.9	30.5	186.9	23.0	187.8
Average	**19.7**	**179.5**	**20.2**	**177.5**	**25.6**	**207.5**	**21.8**	**176.8**

Table 1. MPJPE before and after our appearance-based attacks. We report the results of three networks and observe that disentangled networks are vulnerable to our attacks.

Metric	NSD	DRNet	CSSL
SSIM↑	0.947	0.963	0.943
PSNR↑	24.65	26.45	24.37
MSE↓	0.012	0.007	0.013

Table 2. Quantitative comparison of adversarial images with the original synthesized images. These number show that the images obtained by attacking DRNet are closer to the original synthesized ones, and thus that the DRNet pose vectors tend to contain more appearance information.

6 Discussion

Evaluating Disentanglement. Several methods [8, 6, 19] have been proposed for assessing the degree of disentanglement of latent variables. In particular, we report the two complementary state-of-the-art metrics of [19], Distance Correlation (DC) and Information over Bias (IOB) to evaluate disentanglement. DC is bounded in [0,1] and measures the correlation between the two latent spaces; IoB measures the amount of information from the input image that is encoded in a given latent space. In Table 3, we provide these metrics, averaged over 400 images, for the pose (P) and appearance (A) latent spaces and for different disentanglement strategies. DC(A, P) contain large values indicating that the appearance and pose are correlated. Furthermore, the IOB(I, P) values are larger than the IOB(I, A), which suggests that the pose code encodes more input information than the appearance code. Note that DC(A, P) cannot be used as a standalone metric to interpret disentanglement because low values of DC can also indicate noise in one latent space. While DRNet achieves the best DC(A, P) score, its value of 0.90 IOB(I, A) suggests that the appearance code encodes minimal information. Although these metrics quantify disentanglement, they offer little understanding of the disentanglement issues, and IOB is difficult to

interpret because it is unbounded and requires training an external decoder network whose optimal architecture is unknown. By contrast, our analyses enable a finer-grain understanding of the pose and appearance latent spaces of representation learning strategies for human pose estimation, and provide visual results that are easier to interpret.

Metric	NSD	DRNet	CSSL
DC(A, P)↓	0.88	0.59	0.77
IOB(I, A)↑	0.79	0.90	0.95
IOB(I, P)↑	1.15	1.08	1.29

Table 3. Disentanglement-related metrics for the pose (P) and appearance (A) latent spaces extracted from an input image (I).

Metric	CSSL	CSSL(DA)
SSIM↑	0.943	0.926
PSNR↑	24.37	22.90
MSE↓	0.013	0.018

Table 4. Quantitative comparison of adversarial images with original synthesized images. The images obtained with DA are less similar to original synthesized ones.

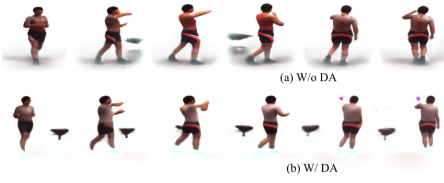


Fig. 8. Synthesizing novel images with CSSL (DA). As in Figure 2, we take S7's pose vector and S8's appearance one and synthesize novel images with CSSL, either without (top) or with (bottom) DA during training. The image synthesized with CSSL (DA) retain S8's appearance without residual red shirt color from S7.

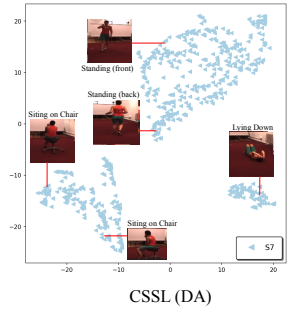


Fig. 9. tSNE visualization of CSSL (DA) appearance codes. The appearance codes of images from same subject are stilll clustered according to the action performed by the subject.

Does data-augmentation help to learn appearance-invariant features? Recently, powerful data augmentation (DA) strategies, such as AugMix [11], CutMix [39] and others [13, 40], have been proposed to improve the generalization power and robustness of neural networks. Furthermore, classical adversarial training [20, 17] can be viewed as a form of data augmentation with adversarial images. Here, we therefore study if data augmentation constitutes a promising direction towards more effectively disentangling self-supervised 3D human pose estimation networks.

Fig. 10. Zero appearance vectors with CSSL (DA). In first row, we show the original image synthesized with CSSL. While, without DA (middle), the synthesized images obtained with a zero appearance vector retain the original subject's appearance, with DA (bottom), all the subjects have a similar the appearance. This suggests that DA helps to remove appearance information from the pose vectors.

Since the network architectures we consider are much more complicated than the image recognition ones used in the above-mentioned DA works, we employ a simpler DA strategy consisting of augmenting the output of the spatial transformer with RGB jitter. We then re-run the analyses we presented before, focusing here on CSSL. Specifically, in Figure 8, we show the images synthesized when mixing S7's pose vector with S8's appearance. Note that, with DA, the images better retain the appearance of S8. Furthermore, in Figure 10, we show images obtained by making use of a zero appearance vector. With DA, all the synthesized images depict a similar subject appearance. Altogether, this suggests that DA helps the disentanglement process in CSSL, which is further confirmed by the DC(A, P) value that improves from 0.77 to 0.62. This value of 0.62 nonetheless still indicates a relatively high correlation between the latent spaces. To further analyze this, we computed a similar t-SNE plot as that of Figure 9, and observed that the actions are still clustered, evidencing that the appearance code still contains some pose information.

Similarly, we also ran our appearance-only attacks on the CSSL model trained with DA, and observed the attacks to remain successful, suggesting that the pose vector remains contaminated by appearance information. To evaluate quantitatively whether DA nonetheless improved this, we report the PSNR, SSIM, and MSE metrics between the attacked images and the original synthesized ones in Table 4. The values indicate that the images obtained by attacking the network without DA are more similar to the original synthesized ones. In other words, CSSL (DA) requires larger changes in the input image to attack the 3D pose regressor. Altogether, these results indicate that DA constitutes a promising direction to improve disentanglement, and we leave the development of more effective DA strategies as future work.

7 Conclusion

In this work, we have analyzed the latent vectors extracted by self-supervised disentangled networks for 3D human pose estimation. Specifically, we have studied the disentanglement of pose and appearance from the perspective of both

the representation learning network, and the supervised 3D human pose regressor. In the former case, our analyses via diverse image synthesis strategies have evidenced that the state-of-the-art disentanglement-based representation learning networks do not truly disentangle pose from appearance, and in particular that the latent pose codes contain significant appearance information. In the latter, we have shown that disentanglement-based networks were not robust to appearance-only adversarial attacks, despite these attacks being designed to be as favorable as possible to the disentanglement-based frameworks. We believe that our analysis methodology and our semantic attacks will be beneficial to improve disentanglement-based representation learning in the future, and thus positively impact self-supervised 3D human pose estimation.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. PMLR (2017)
2. Bhattad, A., Chong, M.J., Liang, K., Li, B., Forsyth, D.: Unrestricted adversarial examples via semantic manipulation. In: International Conference on Learning Representations (2019)
3. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. IEEE (2017)
4. Croce, F., Hein, M.: Sparse and imperceivable adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4724–4732 (2019)
5. Denton, E.L., Birodkar, V.: Unsupervised learning of disentangled representations from video. In: NIPS (2017)
6. Do, K., Tran, T.: Theory and evaluation metrics for learning disentangled representations. In: International Conference on Learning Representations (2019)
7. Dong, Y., Pang, T., Su, H., Zhu, J.: Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4312–4321 (2019)
8. Eastwood, C., Williams, C.K.: A framework for the quantitative evaluation of disentangled representations. In: 6th International Conference on Learning Representations (2018)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
10. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
11. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple data processing method to improve robustness and uncertainty. In: International Conference on Learning Representations (2019)
12. Honari, S., Constantin, V., Rhodin, H., Salzmann, M., Fua, P.: Unsupervised learning on monocular videos for 3d human pose estimation. arXiv preprint arXiv:2012.01511 (2020)
13. Hong, M., Choi, J., Kim, G.: Stylemix: Separating content and style for enhanced data augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14862–14870 (2021)

14. Ionescu, C., Carreira, J., Sminchisescu, C.: Iterated second-order label sensitive pooling for 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1661–1668 (2014)
15. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3334–3342 (2015)
16. Joshi, A., Mukherjee, A., Sarkar, S., Hegde, C.: Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4773–4783 (2019)
17. Kannan, H., Kurakin, A., Goodfellow, I.: Adversarial logit pairing. arXiv preprint arXiv:1803.06373 (2018)
18. Kurakin, A., Goodfellow, I., Bengio, S., et al.: Adversarial examples in the physical world (2016)
19. Liu, X., Thermos, S., Valvano, G., Chartsias, A., O'Neil, A., Tsaftaris, S.A.: Measuring the biases and effectiveness of content-style disentanglement. In: Proceedings of the British Media for VIsion Conference (2021)
20. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018)
21. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2640–2649 (2017)
22. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 international conference on 3D vision (3DV). pp. 506–516. IEEE (2017)
23. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. ACM Transactions on Graphics (TOG) **36**(4), 1–14 (2017)
24. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
25. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2574–2582 (2016)
26. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European symposium on security and privacy (EuroS&P). pp. 372–387. IEEE (2016)
27. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7025–7034 (2017)
28. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Harvesting multiple views for marker-less 3d human pose annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6988–6997 (2017)
29. Popa, A.I., Zanfir, M., Sminchisescu, C.: Deep multitask architecture for integrated 2d and 3d human sensing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6289–6298 (2017)
30. Qiu, H., Xiao, C., Yang, L., Yan, X., Lee, H., Li, B.: Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In: European Conference on Computer Vision. pp. 19–37. Springer (2020)

31. Rhodin, H., Constantin, V., Katircioglu, I., Salzmann, M., Fua, P.: Neural scene decomposition for multi-person motion capture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7703–7713 (2019)
32. Rhodin, H., Salzmann, M., Fua, P.: Unsupervised geometry-aware representation for 3d human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 750–767 (2018)
33. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net: Localization-classification-regression for human pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3433–3441 (2017)
34. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: Adversarial generative nets: Neural network attacks on state-of-the-art face recognition. arXiv preprint arXiv:1801.00349 **2**(3) (2017)
35. Song, Y., Shu, R., Kushman, N., Ermon, S.: Constructing unrestricted adversarial examples with generative models
36. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
37. Tekin, B., Márquez-Neila, P., Salzmann, M., Fua, P.: Learning to fuse 2d and 3d image cues for monocular body pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3941–3950 (2017)
38. Tome, D., Russell, C., Agapito, L.: Lifting from the deep: Convolutional 3d pose estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2500–2509 (2017)
39. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6023–6032 (2019)
40. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018)
41. Zhao, Z., Dua, D., Singh, S.: Generating natural adversarial examples. In: International Conference on Learning Representations (2018)