Semi-Supervised Learning of Disentangled Representations for Cross-Modal Translation

Zheng-Ning Liu
Tsinghua University
Beijing, China
lzhengning@gmail.com

Yan-Pei Cao
Tsinghua University
Beijing, China
caoyanpei@gmail.com

Yu-Tao Yuan
Tsinghua University
Beijing, China
yuanyt17@mails.tsinghua.edu.cn

Tai-Jiang Mu
Tsinghua University
Beijing, China
taijiang@tsinghua.edu.cn

Shi-Min Hu
Tsinghua University
Beijing, China
shimin@tsinghua.edu.cn

Abstract

Collecting paired data for supervised learning can be difficult and expensive, impeding the performance and the generalization ability of trained models. In this paper, we propose a semi-supervised learning approach to cross-modal translation tasks that fully exploits extra data from the target domain. To this end, we employ two interacting variational autoencoders (VAEs) to learn a disentangled representation of the source modal. The first VAE encodes input samples into two disentangled latent spaces, i.e., a task-relevant space and a task-irrelevant space. The task-relevant space is shared by the second VAE that learns to capture the target domain's structure. We then propose a training strategy to guarantee the alignment of the shared latent space, which is the key to high-fidelity translation results. We demonstrate the effectiveness of the proposed method on the task of image-to-sketch translation and 3D human pose estimation from RGB images. Experiment results show that the translation quality of the proposed method achieves state-of-the-art.

1. Introduction

Over the past decades, supervised deep learning techniques have witnessed significant progress in many areas [34, 32, 20, 15] with a large amount of annotated data. To alleviate the laborious labeling demands for some tasks, such as semantic segmentation, text summarization, emotion computation, and 3D human pose estimation, researchers adopt semi-supervised learning algorithms to train their models with a mixture of labeled and unlabeled data. [4, 35, 46, 42, 45, 1, 51].

The labeling process can even be impractical for some cross-modal translation tasks [1, 55] to annotate the desired target for a given source sample. For instance, in the task of image style transfer, the acquisition of pairs of realistic photos and stylized ones requires artistic insights, resulting very high labeling costs. Consequently, the amount of available annotated data is much smaller than in other vision tasks. Another example is the task of 3D human pose estimation from 2D images. Accurate 3D human pose annotations usually can only be reliably collected in multi-camera studios, leading to the lack of variety of human appearances and backgrounds. On the other hand, data in source and target domains can be separately and massively obtained. For example, vast amounts of images and videos containing human actions can be collected from the Internet, while various 3D human poses can be captured using motion capture setups. In this setting, the distribution of the source and target domain can be well characterized separately, and the critical problem of cross-modal translation lies in how to learn a mapping from the source to the target domain with only a relatively small number of annotated data available.

In this paper, we present a flexible method to use the latent structure of the target domain as prior knowledge for cross-modal translation tasks. Inspired by previous works on learning disentangled representations [3, 42, 14, 48, 25, 11, 13, 52], we assume that data in the source domain can be encoded into two disentangled and orthogonal latent spaces, i.e., a *task-relevant* space and a *task-irrelevant* space. Corresponding samples in the target domain can then be generated from the task-relevant latent representations of inputs from the source domain. Taking the task of 3D human pose estimation from RGB images as an example, *poses* and *appearances* of human bodies can be modeled as the task-relevant and task-irrelevant factors, respectively.

With the proposed disentanglement representation, extra data from the target domain can be jointly embedded into

the task-relevant space as well. Therefore, we can bridge the gap between the paired and unpaired data with semi-supervised training. A shared decoder is then used to map the shared task-relevant representation to the target modal. The decoder can guarantee the fidelity of output as it is trained to conform to the target domain's distribution.

To achieve the disentanglement, we employ two interacting variational autoencoders (VAEs) [31] to learn the latent representations. The first VAE encodes samples from the source domain into task-relevant and task-irrelevant spaces. The task-relevant space is shared by the second VAE, which captures the structure of the target domain. We then propose a training strategy to guarantee the alignment of the shared task-relevant latent space by enforcing that samples from both the source and target domains can be accurately reconstructed. During inference, given a sample from the source domain, we first generate its latent representation using the first VAE, then decode the task-relevant code to its counterpart in the target domain using the second VAE. We validate the proposed semi-supervised learning framework on two cross-modal translation tasks, i.e., image-to-sketch translation and 3D human pose estimation from RGB images. Experiment results show that the proposed approach achieves state-of-the-art performance using a limited amount of paired data and generalizes well on unseen inputs.

In summary, we make the following contributions:

- We present a novel semi-supervised learning approach that leverages two interacting VAEs to exploit extra data from the target domain, and effectively learns the disentangled representation for cross-modal translation.

- We conduct thorough qualitative and quantitative experiments to demonstrate that besides the capability of learning disentangled and smooth latent spaces, the proposed learning framework can achieve state-of-the-art performance on tasks such as image-to-sketch translation and 3D human pose estimation.

2. Related Work

Disentangled Representations. Although there is not yet a standard formal definition for disentangled representations [21], they have been widely employed to characterize the generative factors of the underlying input data [3, 41, 42]. InfoGAN [12] learned the disentangled representation by maximizing the mutual information between the observation and a subset of the latent variables based on Generative Adversarial Nets (GANs) [18]. For the task of pose-invariant face recognition, Tran et al. [48] exploited a GAN to explicitly disentangle the identification representation from the pose variation by providing a pose code for the generator to transform a face of the same identity to the target pose and training the discriminator to classify both the identity and the pose. Jiang et al. [25] proposed an encoder-decoder decomposition network with spectral graph convolution to split a 3D face mesh into identity part and expression part.

VAE [31] is a deep generative model that encodes the input data into a latent space and then decodes the latent representation to reconstruct the input data as faithfully as possible. The latent representation can be disentangled into different dimensions or groups to reflect some inherent aspects of the input data. β-VAE [22, 6] is a modification of the original VAE with an adjustable hyperparameter β on the Kullback-Leibler divergence, balancing the channel capacity and independence of the latent representation. Chen et al. [11] introduced a measure on the total correlation between latent variables to encourage the disentanglement, serving as a plug-in replacement for the β-VAE without additional hyperparameters. Kim and Mnih [29] also improved the β-VAE with the total correlation penalty, using an additional discriminator to distinguish the input. Siddharth et al. [42] incorporated a general graphical model into the encoder and decoder of standard VAE to learn a subset of interpretable variables for semi-supervised learning. The VAE architecture is also adopted to learn the disentangled latent features, consisting of time-invariant (e.g., objects) and time-dependent (e.g., dynamics of objects) parts, for sequential data, such as video and audio [16, 37].

All these methods focused on better separation among latent representation of data from the same domain; our approach instead aims for task-specific disentanglement with unpaired data from the target domain.

Cross-Modal Translation. Cross-modal translation has become one of the most popular applications of generative models. Transferring information between modalities with conditional GANs [40, 24, 36, 1, 55] has been applied in various image generation tasks, where adversarial network prompts the distribution of the generator's output close to the target modality. In the 3D pose estimation task, Spurr et al. [47] learn a unified latent space for RGB images and 3D hand pose via two VAEs. Yang et al. [52] propose a two-step training strategy to disentangle the latent factors of multiple modalities for image synthesis and hand pose estimation. Different from the above methods, our method focuses on improving decoders' performance with additional data from the target domain. Gu et al. [19] propose two parallel VAEs and discriminate loss applied for latent space disentanglement. Instead of using GAN to align the latent space explicitly, we exploit the correlation of two spaces with paired data to avoid potential mode collapse and failure to converge.

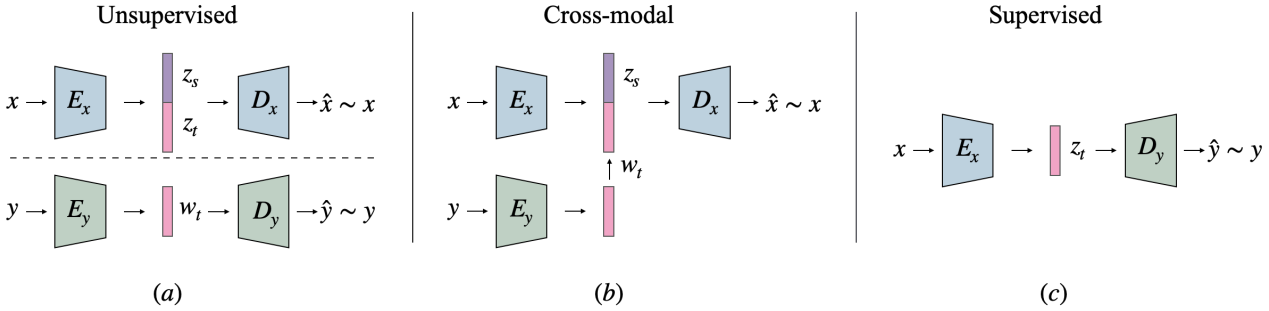

Figure 1. Illustration of our learning objectives. Our framework consists of two interacting VAEs. VAE$_\mathbf{x}$ learns to reconstruct $\mathbf{x}$ from the source modal by encoding them with two disentangled latent variable $\mathbf{z}_t$ and $\mathbf{z}_s$. VAE$_\mathbf{y}$ encodes $\mathbf{y}$ from the target modal into $\mathbf{w}_t$, which shares the same latent space with $\mathbf{z}_t$. The learning objectives are: (a) two vanilla VAE losses that reconstruct the input; (b) a cross-modal reconstruction loss that enforces the alignment of latent spaces between $\mathbf{z}_t$ and $\mathbf{w}_t$; (c) a supervised loss that maximize $p(\mathbf{y}|\mathbf{x})$. $\sim$ stands for loss functions between two variables.

3. Approach

Given N supervised data points of inputs and targets $\mathcal{D}^{sup} = \{(\mathbf{x}^1, \mathbf{y}^1), \ldots, (\mathbf{x}^N, \mathbf{y}^N)\}$, along with additional M unsupervised target data points $\mathcal{D}^t = \{\mathbf{y}^{N+1}, \ldots, \mathbf{y}^{N+M}\}$, our goal is to train a model with parameters Θ that maximize the posterior probability $p_\Theta(\mathbf{y}|\mathbf{x})$. In addition, we suppose that $\mathbf{y}$ can be indirectly inferred from $\mathbf{x}$, and the distribution of $\mathbf{y}$ can be fully characterized with the extra target dataset $\mathcal{D}^t$.

Considering that $\mathbf{y}^i$ from $\mathcal{D}^{sup}$ and $\mathbf{y}^j$ from $\mathcal{D}^t$ are independent and identically distributed, the basic motivation of our approach is using one shared decoder to establish a mapping from shared task-relevant latent codes $\mathbf{z}$ to $\mathbf{y}$. The benefit of this design is two-fold. Firstly, with data from both $\mathcal{D}^{sup}$ and $\mathcal{D}^t$, the trained model has the potential to generalize better than the one trained with $\mathbf{y}$ only from $\mathcal{D}^{sup}$. More importantly, since the decoder is trained to conform to the distribution of the target domain, it can guarantee that the generated outputs retain the target domain's characteristics.

To train the decoder with two sources of data jointly, we should construct two encoders of mapping $\mathbf{x} \to \mathbf{z}$ and mapping $\mathbf{y} \to \mathbf{z}$. There is a large body of existing approaches to learning a latent representation of a single modality mapping $\mathbf{y} \to \mathbf{z} \to \mathbf{y}$, e.g., using vanilla VAEs [31]. However, given only a limited amount of paired data, it can be hard to build a generalizable model that learns the cross-modal mapping from $\mathbf{x}$ to $\mathbf{y}$. In other words, there could be too much information within of the input modality to be distinguished, leading to entanglement and over-fitting. For instance, a convolutional neural network (CNN) might focus on texture details rather than object shapes to classify images [17]. Therefore, we wish to align the task-relevant subspaces of $\mathbf{x}$'s latent representation with the latent space of $\mathbf{y}$ through disentangling $\mathbf{x}$'s latent space.

We propose a semi-supervised learning framework that employs two interacting VAEs to learn the disentangled representation of $\mathbf{x}$. The first VAE, VAE$_\mathbf{x}$, aims to reconstruct $\mathbf{x}$. The difference between a vanilla VAE and VAE$_\mathbf{x}$ is that VAE$_\mathbf{x}$ encodes an input sample into two disentangled latent codes, i.e., a task-relevant $\mathbf{z}_t$ and a task-irrelevant $\mathbf{z}_s$. In particular, $\mathbf{z}_t$ contains the essential information to decode $\mathbf{y}$ without redundancy. The space of $\mathbf{z}_s$ is orthogonal to the space of $\mathbf{z}_t$, which has the complementary information to code $\mathbf{x}$. Another network, namely VAE$_\mathbf{y}$, learns to reconstruct the target $\mathbf{y}$ from the latent representation $\mathbf{w}_t$. $\mathbf{w}_t$ is naturally task-relevant. We choose VAE as our basic architecture, because the latent space of VAE is continuous and exploitable, which is favorable to align the latent space by manipulating latent variables (more details in Section 3.2).

3.1. Train VAEs Separately

Before we introduce the interaction between the two VAEs, we first train each VAE separately. The objective of a VAE is to maximize the log-likelihood of observed samples $\log p(\mathbf{x})$. Directly optimizing $\log p(\mathbf{x})$ is difficult, so we instead maximize the *evidence lower bound* (ELBO) via a latent variable $\mathbf{z}$,

$$\begin{aligned}\log p(\mathbf{x}) &\geq \text{ELBO}(\mathbf{x}, \theta, \phi) \\ &= -\mathbb{E}_{\mathbf{z}\sim q_\phi} \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}, \mathbf{x})} \\ &= \mathbb{E}_{\mathbf{z}\sim q_\phi} \log p_\theta(\mathbf{x}|\mathbf{z}) - \mathbb{D}_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})))\end{aligned} \tag{1}$$

where θ is the parameters of the decoder, and $q_\phi(\mathbf{z}|\mathbf{x})$, usually called the *recognition model*, is parametered by ϕ, focusing on approximating the prior distribution $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. $\mathbb{D}_{KL}(\cdot)$ is the Kullback-Leibler divergence.

In our VAE framework, the objectives of the two VAEs

are slightly modified versions of Eq. 1,

$$\begin{aligned}\mathcal{L}_{\text{VAE}_\mathbf{x}} = &\sum_{i=1}^{N} \mathbb{E}_{\mathbf{z}_t,\mathbf{z}_s\sim q_{\phi_\mathbf{x}}} \log p_{\theta_\mathbf{x}}(\mathbf{x}^i|\mathbf{z}_t,\mathbf{z}_s) \\ &- \beta_\mathbf{x}\mathbb{D}_{KL}(q_{\phi_\mathbf{x}}(\mathbf{z}_t|\mathbf{x}^i)\|p(\mathbf{z}_t)) \\ &- \beta_\mathbf{x}\mathbb{D}_{KL}(q_{\phi_\mathbf{x}}(\mathbf{z}_s|\mathbf{x}^i)\|p(\mathbf{z}_s)) \qquad (2)\\ \mathcal{L}_{\text{VAE}_\mathbf{y}} = &\sum_{j=1}^{N+M} \mathbb{E}_{\mathbf{w}_t\sim q_{\phi_\mathbf{y}}} \log p_{\theta_\mathbf{y}}(\mathbf{y}^j|\mathbf{w}_t) \\ &- \beta_\mathbf{y}\mathbb{D}_{KL}(q_{\phi_\mathbf{y}}(\mathbf{w}_t|\mathbf{y}^j)\|p(\mathbf{w}_t)) \qquad (3)\end{aligned}$$

Here, $(\mathbf{x}_t, \mathbf{x}_s)$ is the disentangled representation for input data and $\mathbf{w}_t$ is the latent representation for the target data. $\theta_\mathbf{x}$, $\theta_\mathbf{y}$, $\phi_\mathbf{x}$, and $\phi_\mathbf{y}$ are the parameters of generation models and recognition models, respectively. In practise, we add auxiliary hyper-parameter $\beta_\mathbf{x}$ and $\beta_\mathbf{y}$ to balance the reconstruction quality and the KL divergence as in [22, 6].

In Eq. 2, the generation process of $\text{VAE}_\mathbf{x}$ is controlled by two latent variables, so there are two KL divergence terms. To train $\text{VAE}_\mathbf{y}$, we employ both $\mathcal{D}^{sup}$ and $\mathcal{D}^t$. Leveraging larger amount of $\mathbf{y}$ promotes $\text{VAE}_\mathbf{y}$'s ability to construct a broader latent space to describe

3.2. Disentangle the latent space

In general, explicitly separating the latent space does not always not contribute to the disentanglement of $\mathbf{z}_t$ and $\mathbf{z}_s$. Therefore, we are facing two challenges: how to achieve the disentanglement and how to guarantee the alignment of $\mathbf{z}_t$ and $\mathbf{w}_t$.

To this end, we propose a cross-modal reconstruction strategy that reconstructs the $\mathbf{x}^i$ from $\{\mathbf{w}_t, \mathbf{z}_s\}$ instead of $\{\mathbf{z}_t, \mathbf{z}_s\}$. And the corresponding loss function is

$$\begin{aligned}\mathcal{L}_{cross} = &\sum_{i=1}^{N} \mathbb{E}_{\mathbf{w}_t\sim q_{\phi_\mathbf{y}},\mathbf{z}_s\sim q_{\phi_\mathbf{x}}} \log p_{\theta_\mathbf{x}}(\mathbf{x}^i|\mathbf{w}_t,\mathbf{z}_s) \\ &- \beta_\mathbf{y}\mathbb{D}_{KL}(q_{\phi_\mathbf{y}}(\mathbf{w}_t|\mathbf{y}^i)\|p(\mathbf{w}_t)) \\ &- \beta_\mathbf{x}\mathbb{D}_{KL}(q_{\phi_\mathbf{x}}(\mathbf{z}_s|\mathbf{x}^i)\|p(\mathbf{z}_s)). \qquad (4)\end{aligned}$$

Specifically, $\mathcal{L}_{cross}$ enforces the $\mathbf{z}_t$ and $\mathbf{w}_t$ to share the same representation, so that the decoder of $\text{VAE}_\mathbf{x}$ can reconstruct $\mathbf{x}$ indistinguishably. At the same time, since $\mathbf{w}_t$ (and thus $\mathbf{z}_t$) fully encodes the characteristics of the target domain, the remaining information of $\mathbf{x}$ encoded in $\mathbf{z}_s$ is then naturally disentangled from $\mathbf{z}_t$.

Given that the shared latent space is well aligned, for an arbitrary sample pair $\{\mathbf{x}^i, \mathbf{y}^i\}$, we can further assume that the posterior distribution of $\mathbf{z}_t$ and $\mathbf{w}_t$ conforms to each other:

$$q_{\theta_\mathbf{x}}(\mathbf{z}_t|\mathbf{x}^i) \approx q_{\theta_\mathbf{y}}(\mathbf{w}_t|\mathbf{y}^i). \qquad (5)$$

With this assumption, we can adopt $\mathbf{z}_t$ to take the place of $\mathbf{w}_t$ in Eq. 3:

$$\begin{aligned}\mathcal{L}_{sup} = &\sum_{i=1}^{N} \mathbb{E}_{\mathbf{z}_t\sim q_{\phi_\mathbf{x}}} \log p_{\theta_\mathbf{y}}(\mathbf{y}^i|\mathbf{z}_t) \\ &- \beta_\mathbf{x}\mathbb{D}_{KL}(q_{\phi_\mathbf{x}}(\mathbf{z}_t|\mathbf{x}^i)\|p(\mathbf{z}_t)). \qquad (6)\end{aligned}$$

Here, $\mathcal{L}_{sup}$ plays the role of a supervised loss that maximize $p(\mathbf{y}^i|\mathbf{x}^i)$ with the normal restriction of latent code $\mathbf{z}_t$. In fact, even if the assumption Eq. 5 does not hold perfectly, the supervised loss will push $\mathbf{z}_t$ close to $\mathbf{w}_t$ since the decoding outputs have to be the same. Consequently, $\mathcal{L}_{sup}$ also imposes the ability of pushing task-relevant information into $\mathbf{z}_t$, and thus strengthens the disentanglement. Note that both $\mathcal{L}_{cross}$ and $\mathcal{L}_{sup}$ are defined with paired training data.

The above process also applies to the inference stage. To be more clear, to translate $\mathbf{x}$ to $\mathbf{y}$, we first send $\mathbf{x}$ into $\text{VAE}_\mathbf{x}$'s encoder to obtain $\mathbf{z}_t$ and then decode $\mathbf{z}_t$ with $\text{VAE}_\mathbf{y}$'s decoder.

3.3. Training an End-to-End Model

We then train $\text{VAE}_\mathbf{x}$ and $\text{VAE}_\mathbf{y}$ jointly in an end-to-end manner. The overall objective is:

$$\mathcal{L} = \mathcal{L}_{\text{VAE}_\mathbf{x}} + \mathcal{L}_{\text{VAE}_\mathbf{y}} + \mathcal{L}_{cross} + \mathcal{L}_{sup}. \qquad (7)$$

Usually, the scale of target dataset M is much larger than N. To make full use of M, we add $m = \lfloor \frac{\|M\|}{\|N\|} \rfloor$ steps to train $\text{VAE}_\mathbf{y}$ with unpaired target samples after each iteration of joint training. The overall framework of our method is presented in Algorithm 1.

4. Experiments

We evaluate our framework on two computer vision tasks to demonstrate the disentanglement ability of our method and the performance improvement that our method brings about.

4.1. Image-to-Sketch Translation

Image-to-sketch translation aims to establish a mapping from realistic images to human drawn sketches. It is a typical task that owns tremendous unlabeled from each domain but lacks paired annotations. Here, we choose to generate a monochrome sketch image for a photo.

We use the QMUL-Shoe-Chair-V2[53], which is the largest fine-grained image-sketch dataset. The dataset contains 2,000 real photos of shoes, 6,648 sketches, which are split into training and testing sets with a ratio of $9:1$. The sketches are preprocessed with distance transform [5] to facilitate a more stable training process. The network structures of $\text{VAE}_\mathbf{x}$ and $\text{VAE}_\mathbf{y}$ are the same, using ResNet18 [20]

Algorithm 1: Training procedure

initialize $\theta_{\mathbf{x}}$, $\phi_{\mathbf{x}}$, $\theta_{\mathbf{y}}$, $\phi_{\mathbf{y}}$;
repeat
  Sample minibatch of $\mathbf{x}$, $\mathbf{y}$ from the supervised dataset $\mathcal{D}^{sup}$;
  $q_{\phi_{\mathbf{x}}}(\mathbf{z}_t, \mathbf{z}_s|\mathbf{x}) \leftarrow$ encode $\mathbf{x}$;
  $q_{\phi_{\mathbf{y}}}(\mathbf{w}_t|\mathbf{y}) \leftarrow$ encode $\mathbf{y}$;
  $p_{\theta_{\mathbf{x}}}(\mathbf{x}|\mathbf{z}_t, \mathbf{z}_s) \leftarrow$ decode $\mathbf{z}_t, \mathbf{z}_s$;
  $p_{\theta_{\mathbf{y}}}(\mathbf{y}|\mathbf{w}_t) \leftarrow$ decode $\mathbf{w}_t$;
  $p_{\theta_{\mathbf{x}}}(\mathbf{x}|\mathbf{w}_t, \mathbf{z}_s) \leftarrow$ decode $\mathbf{w}_t, \mathbf{z}_s$;
  $p_{\theta_{\mathbf{y}}}(\mathbf{y}|\mathbf{z}_t) \leftarrow$ decode $\mathbf{z}_t$;
  $\mathcal{L} = \mathcal{L}_{\text{VAE}_{\mathbf{x}}} + \mathcal{L}_{\text{VAE}_{\mathbf{y}}} + \mathcal{L}_{cross} + \mathcal{L}_{sup}$, by Eq.2, 3, 4, 6;
  Update $\theta_{\mathbf{x}}$, $\phi_{\mathbf{x}}$, $\theta_{\mathbf{y}}$, $\phi_{\mathbf{y}}$ using Adam [30];
  for $i \leftarrow 1$ **to** m **do**
    Sample minibatch of $\mathbf{y}$ from the target dataset $\mathcal{D}^t$;
    $q_{\phi_{\mathbf{y}}}(\mathbf{w}_t|\mathbf{y}) \leftarrow$ encode $\mathbf{y}$;
    $p_{\theta_{\mathbf{y}}}(\mathbf{y}|\mathbf{w}_t) \leftarrow$ decode $\mathbf{w}_t$;
    $\mathcal{L}_{\text{VAE}_{\mathbf{y}}} \leftarrow$ calculate Eq. 3 with $q_{\phi_{\mathbf{y}}}(\mathbf{w}_t|\mathbf{y})$, $p_{\theta_{\mathbf{y}}}(\mathbf{y}|\mathbf{w}_t)$;
    Update $\theta_{\mathbf{y}}$, $\phi_{\mathbf{y}}$ using Adam;
  end
until *convergence of parameters* ($\theta_{\mathbf{x}}$, $\phi_{\mathbf{x}}$, $\theta_{\mathbf{y}}$, $\phi_{\mathbf{y}}$);

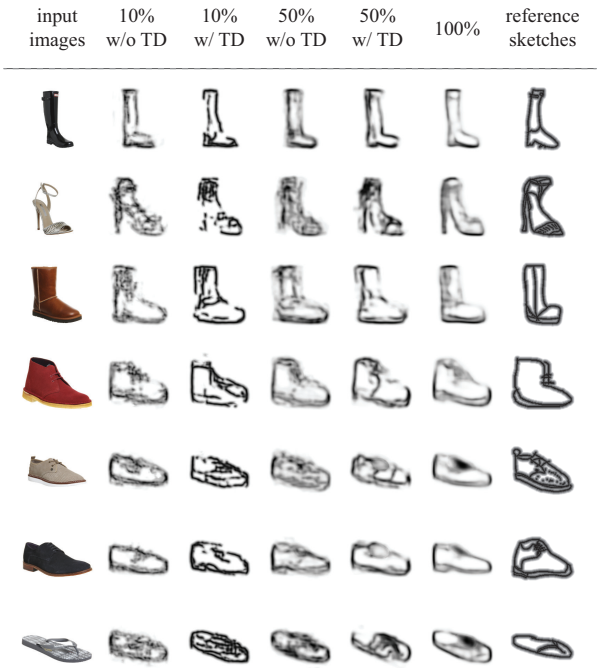

Figure 2. Image-to-sketch translation results with different ratios of supervised data. TD denotes the target dataset which consists of the unused sketches.

Figure 3. Overfitting examples in the training dataset. Images from left to right are input images, reference sketches, and generated sketches. The network is trained with only 10% of the paired data. Sketch patterns of other shoes can be seen in the generated sketches.

Table 1. Comparison of 3D human pose estimation results on Human3.6M. 10% denotes that 10% training samples in Human3.6M are used for training.

Method	MPJPE	PA-MPJPE
AIGNs [49]	-	97.2
Chen et al. [8]	69.1	51.9
Martinez et al. [39]	**62.9**	**47.7**
Pavlakos et al. [44]	71.9	51.9
Zhou et al. [54]	64.9	-
Kundu et al (10%). [43]	-	50.8
Ours (10%)	**68.7**	**49.7**
Ours	**68.1**	**49.2**

as encoders, and the decoders are composed of convolution, BatchNorm, ReLU, and upsampling operations. All latent dimensions are set to 64. We use L2 losses to reconstruct the images and sketches. The Adam optimizer is employed to train the VAEs with a learning rate of $1e-4$. Please see the supplementary material for more details.

Results. We conduct an ablation study of supervised data to evaluate the effectiveness of unpaired data (see Fig. 2). We randomly discard different ratios of real photos and use the left sketches without paired real photos as the unpaired target dataset. The strokes of the third and the fifth columns are much more clear than those from the second and the fourth columns. This highlights that the proposed method does enforce the sketch decoder to conform to the real distribution of shoes sketches.

Also, we notice that the generated sketches are noisier when the model is trained without the extra sketch dataset (see the second and third columns). We find the reason is that the network overfits the training samples (see Fig. 3). When trained with only 10% paired images and sketches, the network easily outputs a local optimum, which is a mixture of multiple sketches, resulting in the noise on the testing set. Therefore, the proposed method is able to prevent overfitting when lacking enough training data.

4.2. 3D Human Pose Estimation from RGB Images

3D human pose estimation aims to understand the human pose in 3D spaces. The acquisition process of RGB images with ground-truth 3D human pose annotations is expensive and can be impractical for in-the-wild scenarios. Nevertheless, obtaining solely the 3D poses is much easier, thanks to

the motion capture system. Formally, given a RGB image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$, the target $\mathbf{y} \in \mathbb{R}^{3\times J}$ is the 3D position of J joints of the person in the image $\mathbf{x}$. In this scenario, $\mathbf{z}_t$ encodes the 3D human pose, and $\mathbf{z}_s$ can be regarded as other factors, e.g., the appearance and background.

Dataset. Human3.6M [23] is a commonly used in-studio dataset for 3D human pose estimation, which consists of 3.6 million video frames varying in 4 camera views, 11 subjects and 15 actions. We follow the previous works [39, 49, 50, 9, 43] to split training and testing set. Specifically, the training set consists of 5 subjects, (S1, S5, S6, S7, S8), and the other 2 subjects, (S9, S11), are used for evaluation. However, the variety of appearances of Human3.6M is not enough to learn a robust visual feature for in-the-wild images, thus we use two additional 2D datasets, i.e., LSP [26, 27] and MPII human pose dataset [2], as a weak supervision by projecting the estimated 3D pose to the 2D image plane with a regressed weak camera model [28]. We adopt the CMU MoCap database [33] from AMASS [38] as the extra target dataset, which has around 9 hours of 3D pose sequences.

Data processing. We crop the input images according to ground-truth bounding boxes so that the human bodies are centered, and then scale the images to 224x224 pixels. When ground-truth 2D annotations are not available, we utilize the 2D joints from OpenPose [7] to estimate the bounding boxes. Furthermore, we apply Deeplab V3 [10] to roughly remove the background as we are only interested in the foreground. For Human3.6M, the provided background segmentation masks are used.

Training details. We use ResNet-50 [20] as the backbone of $\text{VAE}_{\mathbf{x}}$'s encoder to extract visual features. For the decoder, we stack four layers of convolution, BatchNorm, ReLU, and upsampling to reconstruct the images to the original resolution. Both the encoder and decoder of $\text{VAE}_{\mathbf{y}}$ are composed of three repetitions of fully-connected, BatchNorm, and LeakyReLU layers. We use an additional linear layer after the ResNet-50 to regress the weak camera parameters. The dimension of each latent space is 64. L2 losses are used for RGB image reconstruction, joint position reconstruction, and 3D pose estimation. The model is trained using Adam optimizer with a learning rate of $1e-4$ for 50 epoches. Please refer to our supplementary material for more details.

Results. We evaluate our method with both the mean per joint position error (MPJPE) and the MPJPE after Procrustes Alignment (PA-MPJPE), measured in millimeters. The quantitative result is reported in Table 1. AIGNs [49], Chen et al. [8], and Kundu et al. [43] also used unpaired 3D skeleton dataset. AIGNs use the external 3D poses as real samples in a discrimative loss. Chen et al. perform a 2D pose matching in the 3D poses dataset. Similar to our method, Kundu et al. employ unpaired 3D poses to constrain the latent space towards generating real pose distribution with a decoupled energy minimization strategy. In comparison, we achieve this goal through a probabilistic model. In the Human3.6M dataset, our approach is more accurate than the above methods. Our approach also outperforms all other methods but Martinez et al. in terms of PA-MPJPE, indicating that the latent space indeed learns a distribution of natural 3D human poses. Martinez et al. uses a well-trained out-of-the-box 2D joint detector, which is a strong prior for 3D joint estimation. On the contrary, we train the end-to-end network from scratch. The top four rows in Fig. 4 gives some results in the test set of Human3.6M.

Table 2. Ablation study of whether to apply $\mathcal{L}_{cross}$ in the training objective

	w/o $\mathcal{L}_{cross}$	w/ $\mathcal{L}_{cross}$
MPJPE	87.4	68.9
PA-MPJPE	55.0	49.7

Table 3. Comparison of the proposed method and the simple baseline on the subject #1 of Human3.6m. The baseline method only employs the supervision term.

	baseline	full losses
PA-MPJPE	76.7	60.6

Ablation Study. To evaluate the importance of $\mathcal{L}_{cross}$, which enforces Eq. 5, we conduct an experiment where we set the weight of $\mathcal{L}_{cross}$ to zero while other experiment settings remain the same. In Table 2 , the corresponding MPJPE and MPJPE (PA) on the test set are worse than the case of the full loss function. This indicates that without $\mathcal{L}_{cross}$, the misalignment of the latent space can undermine the performance of the shared decoder.

We also compared the proposed framework with a simple baseline, which has only 2D and 3D supervisions. In particular, we set the overall loss to be $\mathcal{L} = \mathcal{L}_{sup}$. We conduct the experiment on the first sequence (S1) of the Human3.6M dataset. Other experiment settings are kept unchanged. Table 3 shows the result. This experiment shows the effectiveness of the additional target dataset and the proposed framework.

In order to assess the effectiveness of the unpaired poses in 3D human pose estimation, we evaluate our framework with different amounts of training supervision. We use part of paired images and 3D poses from Human3.6M to train the networks and discard other data. For very small subsets, e.g., 5% and 10% of S1, the images are sampled randomly. Fig. 5 indicates that our approach can achieve a small performance degradation with a significantly decreased amount of 3D annotations.

Disentanglement. We then visualize the learnt latent space by interpolating and manipulating the latent codes. Fig. 6(a) displays that the latent space of $\mathbf{z}_t$ is smooth and

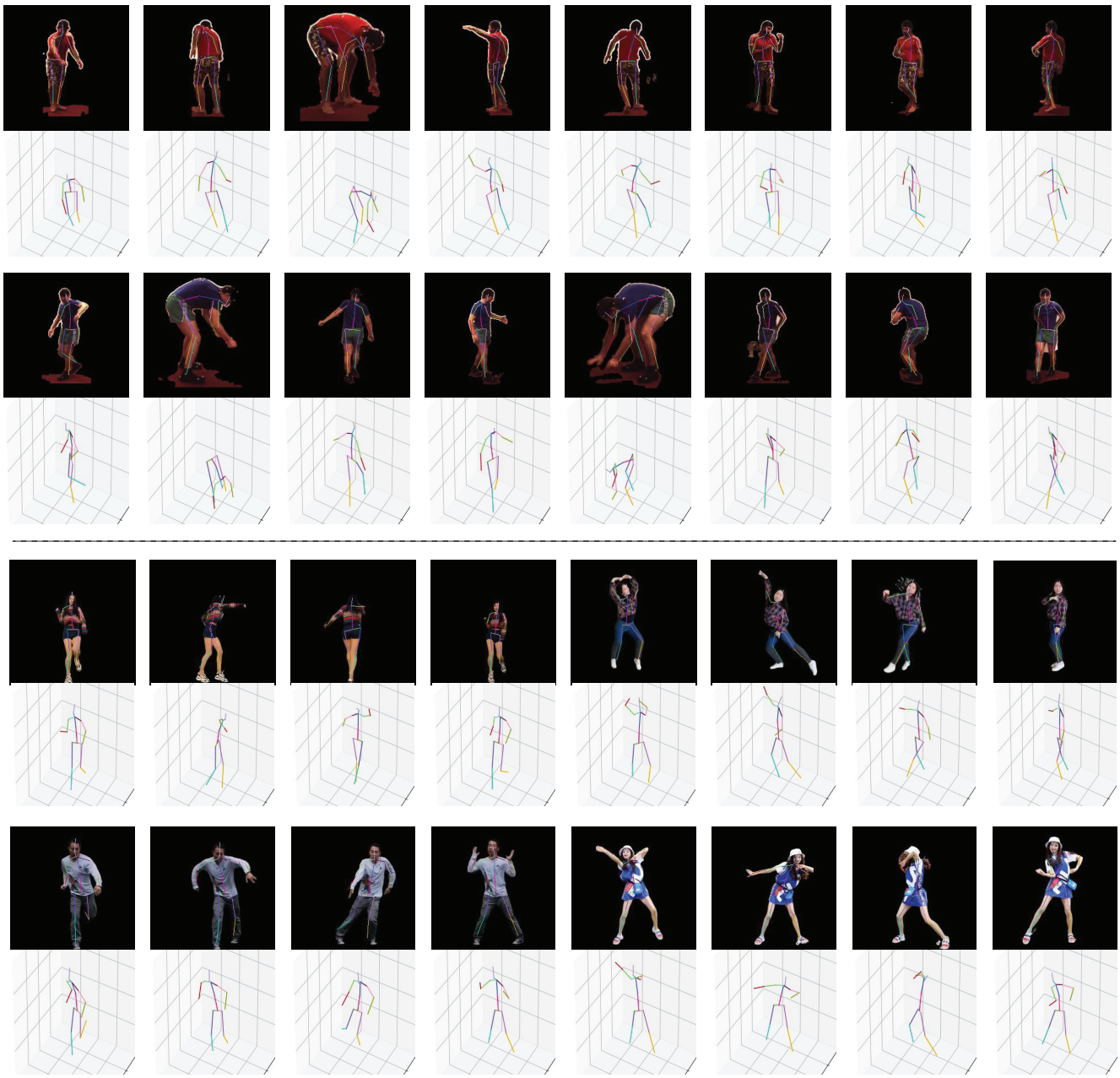

Figure 4. Results gallery of estimated 3D human pose. Top four rows: examples from the Human3.6M test set. Bottom four rows: estimations on dancing videos from the Internet. Skeletons are visualized from another viewing angle under each RGB image overlays. More results are available in the supplementary materials.

consistent, without impossible poses. To evaluate the disentanglement of $\mathbf{x}$'s latent space, we visualize this through the output images of $\text{VAE}_{\mathbf{x}}$. We randomly sample several $\mathbf{z}_t$ and $\mathbf{z}_s$ from the Human3.6M test set. These $\mathbf{z}_t$ and $\mathbf{z}_s$ are then manipulated as combinations to generate images, illustrated in Fig. 6(b). We can clearly see that $\mathbf{z}_t$ controls the human pose, which is task-relevant. Meanwhile, images with the same $\mathbf{z}_s$ almost shares the same clothes, indicating that appearance is task-irrelevant. It also proves that our pose estimation model is robust to the variety of appearances.

5. Conclusions

We have introduced a novel semi-supervised approach to learning disentangled latent representations for cross-modal translation tasks. We show that using extra unpaired data from the target domain helps to disentangle the task-relevant part of the latent representation. The proposed

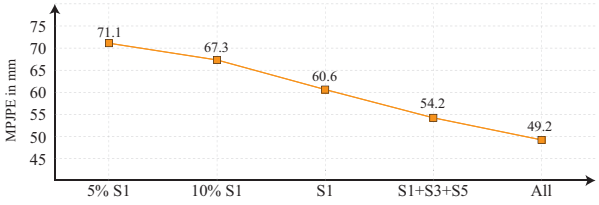

Figure 5. Performances of 3D human pose estimation on Human3.6M as a function of the amount of training supervision.

training objective ensures the alignment of the two latent spaces and then realize the cross-modal translation. Experiment results on the tasks of image-to-sketch translation and 3D human pose estimation demonstrate the generalization ability of the proposed method. The proposed method also achieves the state-of-the-art performance for 3D human pose estimation. In the future, we may consider to extend the proposed framework to make use of unpaired data in the source domain.

References

[1] A. Almahairi, S. Rajeshwar, A. Sordoni, P. Bachman, and A. Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *International Conference on Machine Learning*, pages 195–204, 2018. 1, 2

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 6

[3] Y. Bengio, A. C. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013. 1, 2

[4] K. P. Bennett and A. Demiriz. Semi-supervised support vector machines. In *Advances in Neural Information processing systems*, pages 368–374, 1999. 1

[5] G. Borgefors. Distance transformations in digital images. *Comput. Vis. Graph. Image Process.*, 34(3):344–371, 1986. 4

[6] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner. Understanding disentangling in β-vae. *CoRR*, abs/1804.03599, 2018. 2, 4

[7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 6

[8] C. Chen and D. Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5759–5767, 2017. 5, 6

[9] C. Chen, A. Tyagi, A. Agrawal, D. Drover, R. MV, S. Stojanov, and J. M. Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *CVPR*, pages 5714–5724. Computer Vision Foundation / IEEE, 2019. 6

[10] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 6

[11] T. Q. Chen, X. Li, R. B. Grosse, and D. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*, pages 2615–2625, 2018. 1, 2

[12] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, pages 2172–2180, 2016. 2

[13] R. de Bem, A. Ghosh, T. Ajanthan, O. Miksik, N. Siddharth, and P. Torr. A semi-supervised deep generative model for human body analysis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 1

[14] Z. Deng, H. Zhang, X. Liang, L. Yang, S. Xu, J. Zhu, and E. P. Xing. Structured generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 3899–3909, 2017. 1

[15] R. Fan, M. Cheng, Q. Hou, T. Mu, J. Wang, and S. Hu. S4net: Single stage salient-instance segmentation. In *CVPR*, pages 6103–6112. Computer Vision Foundation / IEEE, 2019. 1

[16] M. Fraccaro, S. Kamronn, U. Paquet, and O. Winther. A disentangled recognition and nonlinear dynamics model for unsupervised learning. In *NIPS*, pages 3601–3610, 2017. 2

[17] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*. OpenReview.net, 2019. 3

[18] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 2

[19] J. Gu, Z. Wang, W. Ouyang, W. Zhang, J. Li, and L. Zhuo. 3d hand pose estimation with disentangled cross-modal latent space. In *WACV*, pages 380–389. IEEE, 2020. 2

[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016. 1, 4, 6

[21] I. Higgins, D. Amos, D. Pfau, S. Racanière, L. Matthey, D. J. Rezende, and A. Lerchner. Towards a definition of disentangled representations. *CoRR*, abs/1812.02230, 2018. 2

[22] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR (Poster)*. OpenReview.net, 2017. 2, 4

[23] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 6

[24] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 5967–5976. IEEE Computer Society, 2017. 2

[25] Z. Jiang, Q. Wu, K. Chen, and J. Zhang. Disentangled representation learning for 3d face shape. In *CVPR*, pages 11957–11966. Computer Vision Foundation / IEEE, 2019. 1, 2

[26] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12. 6

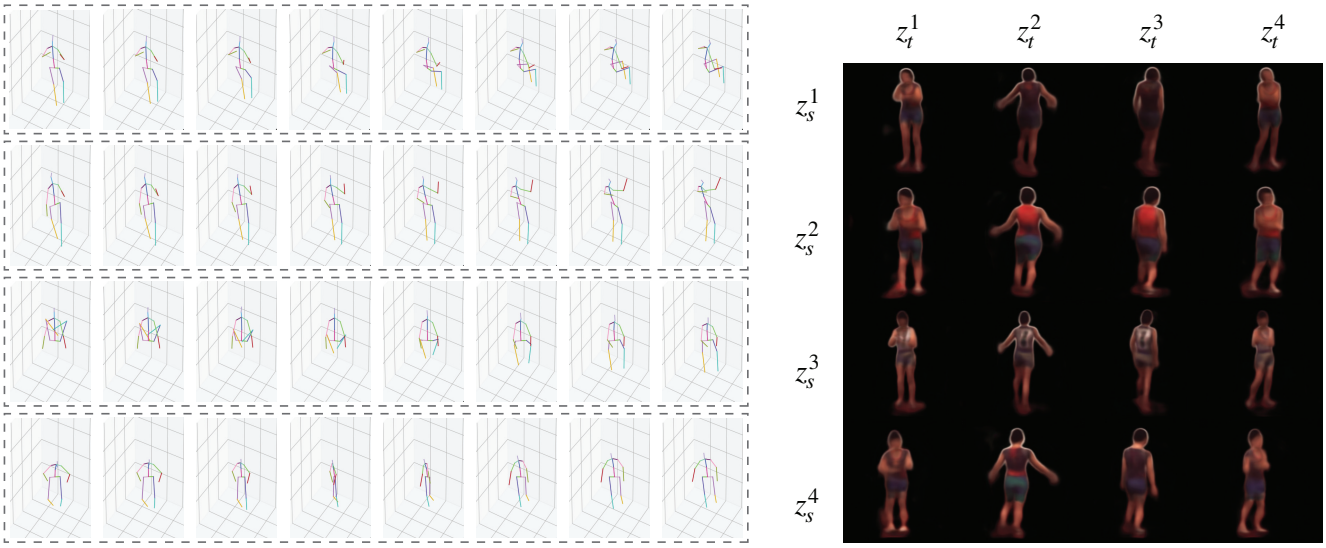

Figure 6. Latent space visualization. (a) Interpolation of $\mathbf{z}_t$'s latent space, where the first column and the last column are samples from the Human3.6M test set. (b) Synthesized images with the combination of 4 $\mathbf{z}_t$'s and 4 $\mathbf{z}_s$'s, which are sampled in the Human3.6M test set. Images in the same column share one $\mathbf{z}_t$ and images in the same row share the same $\mathbf{z}_s$. More results can be found in the supplementary materials.

[27] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 6

[28] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3d human dynamics from video. In *CVPR*, pages 5614–5623. Computer Vision Foundation / IEEE, 2019. 6

[29] H. Kim and A. Mnih. Disentangling by factorising. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 2654–2663. PMLR, 2018. 2

[30] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. 5

[31] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2, 3

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012. 1

[33] C. G. Lab. Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu. 6

[34] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[35] D.-H. Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 2, 2013. 1

[36] H. Lee, H. Tseng, J. Huang, M. Singh, and M. Yang. Diverse image-to-image translation via disentangled representations. In *ECCV (1)*, volume 11205 of *Lecture Notes in Computer Science*, pages 36–52. Springer, 2018. 2

[37] Y. Li and S. Mandt. Disentangled sequential autoencoder. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 5656–5665. PMLR, 2018. 2

[38] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. AMASS: archive of motion capture as surface shapes. In *ICCV*, pages 5441–5450. IEEE, 2019. 6

[39] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, pages 2659–2668. IEEE Computer Society, 2017. 5, 6

[40] M. Mirza and S. Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. 2

[41] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. A. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. 2

[42] S. Narayanaswamy, B. Paige, J. van de Meent, A. Desmaison, N. D. Goodman, P. Kohli, F. D. Wood, and P. H. S. Torr. Learning disentangled representations with semi-supervised deep generative models. In *NIPS*, pages 5925–5935, 2017. 1, 2

[43] J. Nath Kundu, S. Seth, V. Jampani, M. Rakesh, R. Venkatesh Babu, and A. Chakraborty. Self-supervised 3d human pose estimation via part guided novel image synthesis. *arXiv*, pages arXiv–2004, 2020. 5, 6

[44] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7025–7034, 2017. 5

[45] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[46] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko. Semi-supervised learning with ladder networks. In *Advances in neural information processing systems*, pages 3546–3554, 2015. 1

[47] A. Spurr, J. Song, S. Park, and O. Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, pages 89–98. IEEE Computer Society, 2018. 2

[48] L. Tran, X. Yin, and X. Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *CVPR*, pages 1283–1292. IEEE Computer Society, 2017. 1, 2

[49] H. F. Tung, A. W. Harley, W. Seto, and K. Fragkiadaki. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *ICCV*, pages 4364–4372. IEEE Computer Society, 2017. 5, 6

[50] B. Wandt and B. Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *CVPR*, pages 7782–7791. Computer Vision Foundation / IEEE, 2019. 6

[51] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019. 1

[52] L. Yang and A. Yao. Disentangling latent hands for image synthesis and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9877–9886, 2019. 1, 2

[53] Q. Yu, F. Liu, Y.-Z. SonG, T. Xiang, T. Hospedales, and C. C. Loy. Sketch me that shoe. In *Computer Vision and Pattern Recognition*, 2016. 4

[54] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 398–407, 2017. 5

[55] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2242–2251. IEEE Computer Society, 2017. 1, 2