

Regular Paper

MirrorNet: A Deep Reflective Approach to 2D Pose Estimation for Single-Person Images

TAKAYUKI NAKATSUKA^{1,a)} KAZUYOSHI YOSHII² YUKI KOYAMA³ SATORU FUKAYAMA³
 MASATAKA GOTO³ SHIGEO MORISHIMA⁴

Received: August 26, 2020, Accepted: February 2, 2021

Abstract: This paper proposes a statistical approach to 2D pose estimation from human images. The main problems with the standard supervised approach, which is based on a deep recognition (image-to-pose) model, are that it often yields anatomically implausible poses, and its performance is limited by the amount of paired data. To solve these problems, we propose a semi-supervised method that can make effective use of images with and without pose annotations. Specifically, we formulate a hierarchical generative model of poses and images by integrating a deep generative model of poses from pose features with that of images from poses and image features. We then introduce a deep recognition model that infers poses from images. Given images as observed data, these models can be trained jointly in a hierarchical variational autoencoding (image-to-pose-to-feature-to-pose-to-image) manner. The results of experiments show that the proposed reflective architecture makes estimated poses anatomically plausible, and the pose estimation performance is improved by integrating the recognition and generative models and also by feeding non-annotated images.

Keywords: 2D pose estimation, amortized variational inference, variational autoencoder, mirror system

1. Introduction

Human beings understand the essence of things by abstraction and embodiment. As Richard P. Feynman, the famous physicist, stated, “What I cannot create, I do not understand” [15], abstraction and embodiment are two sides of the same coin. Our hypothesis is that such a bidirectional framework plays a key role in the brain process of recognizing human poses from 2D images, inspired by the *mirror neuron system* or *motor theory* known in the field of cognitive neuroscience [17]. In this paper, we focus on the estimation of the 2D pose (joint coordinates) of a person in an image, inspired by the human mirror system.

The standard approach to 2D pose estimation is to train a deep neural network (DNN) that maps an image to a pose in a supervised manner by using a collection of images with pose annotations [3], [30], [42], [44], [45], [48], [51]. Toshev and Szegedy [45] pioneered a method called DeepPose that uses a DNN consisting of convolutional and fully connected layers for the nonlinear regression of 2D joint coordinates from images. Instead of directly using 2D joint coordinates as target data, Thompson et al. [44] proposed a heatmap representation that indicates the posterior distribution of each joint over pixels. This representation has commonly been used in many state-of-the-art methods for 2D pose estimation [3], [30], [42], [48], [51]. Note that all of these methods focus only on the recognition part of the

human mirror system.

Such a supervised approach based on *image-to-pose* mapping has two major drawbacks. First, the anatomical plausibility of estimated poses is not taken into account. To mitigate this problem, the positional relationships between adjacent joints have often been considered [4], [9], [26], [32], [43], and error correction networks [5], [6] and adversarial networks [7], [8] have been used in a heuristic manner. Second, the performance of the supervised approach is limited by the amount of paired pose-image data. To overcome this limitation, data augmentation techniques [34] and the use of metadata [46] and non-annotated data [12], [46] have been proposed. A unified solution to these complementary problems, however, remains an open question.

In this paper, we propose a hierarchical variational autoencoder (VAE) called *MirrorNet* that consists of higher- and lower-level mirror systems (**Fig. 1**). Specifically, we formulate a probabilistic latent variable model that integrates a deep generative model of poses from pose features (called *primitives*) with that of images from poses and foreground and background features (called *appearances* and *scenes*). To estimate poses, pose features, and image features from given images in the framework of amortized variational inference (AVI) [24], we introduce deep recognition models of pose features from poses, foreground and background image features from poses and images, and poses from images. These generative and recognition models can be trained jointly even from non-annotated images.

A key feature of our semi-supervised method is to consider the anatomical *fidelity* and *plausibility* of poses in the estimation process. To make use of both annotated and non-annotated images, our method constructs an *image-to-pose-to-image* reflective model (i.e., a higher-level mirror system for image understand-

¹ Waseda University, Shinjuku, Tokyo 169–8050, Japan

² Kyoto University, Kyoto 606–8501, Japan

³ National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305–8568, Japan

⁴ Waseda Research Institute for Science and Engineering, Shinjuku, Tokyo 169–8050, Japan

^{a)} t59nakatsuka@fuji.waseda.jp

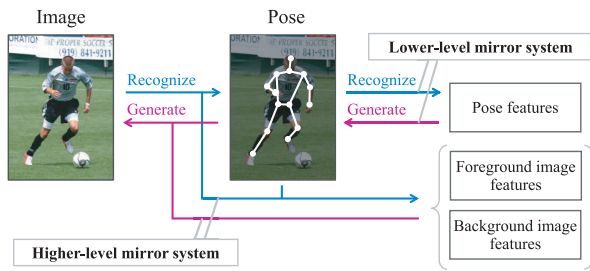


Fig. 1 An overview of MirrorNet, which consists of generative models of poses and images from latent features and recognition models of poses and latent features from images. The latent features consist of primitives (pose features), appearances (foreground image features), and scenes (background image features). A higher-level image-to-pose-to-image mirror system is integrated with a lower-level pose-to-feature-to-pose mirror system in a hierarchical manner for effectively using images with and without pose annotations.

ing) by connecting the *image-to-pose* recognition model with the *pose-to-image* generative model. Even when only images without pose annotations are given, the generative model can be used for evaluating the anatomical fidelity of poses estimated by the recognition model (i.e., how consistent the estimated poses are with the given images). In the same way, our method builds a *pose-to-feature-to-pose* reflective model (i.e., a lower-level mirror system for pose understanding) by connecting the *pose-to-feature* recognition model with the *feature-to-pose* generative model. This pose VAE can be trained in advance by using a large number of pose data (e.g., data obtained by rendering human 3D models) and then used for evaluating the anatomical plausibility of the estimated poses. Note that the pose VAE cannot be used alone as an evaluator of an estimated pose for a non-annotated image because *any* plausible pose is allowed even if it does not reflect the image. This is why conventional plausibility-aware methods still need paired data [6], [7], [8]. The higher- and lower- mirror systems are integrated into MirrorNet and can be trained jointly in a statistically principled manner. In practice, each component of MirrorNet is trained separately by using paired data and then the entire MirrorNet is jointly trained in a semi-supervised manner using both paired and unpaired data for further optimization.

The main contribution of this paper is to realize plausibility- and fidelity-aware 2D pose estimation based on a hierarchical VAE consisting of pose and image VAEs corresponding to the higher- and lower-level mirror systems. To achieve this, we effectively integrate state-of-the-art DNN-based methods for supervised pose estimation, foreground/background segmentation, and image generation into the unified VAE architecture. The VAE-based probabilistic formulation can make effective use of both annotated and non-annotated images for joint semi-supervised learning of pose fidelity and plausibility, leading to improved pose estimation. We experimentally show that the image-to-pose recognition model can be improved by integrating the pose-to-image generative model and the pose VAE.

The rest of this paper is organized as follows. Section 2 reviews related work on plausibility-aware pose estimation and fidelity-aware image processing. Section 3 explains the proposed method for unsupervised, supervised, and semi-supervised pose estimation. Section 4 describes the detailed implementation of the proposed method. Section 5 reports comparative experiments con-

ducted for evaluating the proposed method. Section 6 summarizes this paper.

2. Related Work

Here, 2D human pose estimation refers to estimating the coordinates of joints of a person in an image. This task is challenging because a wide variety of human appearances and background scenes can exist and some joints are often occluded.

For robust pose estimation, Ramanan [37] proposed an edge-based model, and Andriluka et al. [2] introduced a pictorial structural model of human joints. Modeling the human body using tree or graph structures has been intensively studied [11], [14], [19], [35], [40], [41], [52]. To improve the accuracy of estimation, one needs to carefully design sophisticated models and features that can appropriately represent the relations between joints.

Toshev and Szegedy [45] proposed a neural pose recognizer called DeepPose that estimates the positions of joints by using a DNN consisting of convolutional layers and fully connected layers. DeepPose is the first method that applies deep learning to pose estimation, resulting in significant performance improvement. Instead of directly regressing the coordinates of joints from an image as in DeepPose, Thompson et al. [44] used a heatmap (pixel-wise likelihood) for representing the distributions of each joint, which has recently become standard. These state-of-the-art methods for 2D pose estimation have been examined from several points of view. For example, intermediate supervision and multi-stage learning were proposed for using deep convolutional neural networks (CNNs) [3], [30], [42], [48], [51]. An optimal objective function was proposed for evaluating the relations between pairs of joints [4], [9], [26], [43]. Recently, some studies have assessed the correctness of inferred poses using additional networks [6], [13], [29] or compensated for the lack of data samples with data augmentation [34], [46]. Here, we review plausibility-aware methods of pose estimation and fidelity-aware methods of image processing.

2.1 Plausibility-Aware Pose Estimation

A standard way of improving the anatomical plausibility of estimated poses is to focus on the local relations of adjacent joints in pose estimation [4], [9], [26], [32], [43] or to refine the estimated poses as post-processing. Carreira et al. [5] proposed a self-correcting model based on iterative error feedback. Chen et al. [6], Fieraru et al. [13], and Moon et al. [29] proposed cascaded networks that recursively refine the estimated poses while referring to the original images. Adversarial networks have often been used to judge whether the estimated poses are anatomically plausible [7], [8]. In addition, Ke et al. [20] proposed a scale-robust method based on a multi-scale network with a body structure-aware loss function. Nie et al. [31] proposed a structured pose representation using the displacement in the position of every joint from a root joint position. While these methods can use only paired data for supervised learning, our VAE-based method enables unsupervised learning. In contrast to the existing autoencoding approach that aims to extract latent features of poses [27], [47], our VAE is used for measuring the plausibility of poses.

To compensate for a lack of training data, Ukita and Uematsu [46] took a semi- and weakly-supervised approach that uses non-annotated images and action labels of images (e.g., baseball and volleyball) to estimate the poses of humans from a part of paired data. Peng et al. [34] proposed an efficient data augmentation method that generates hard-to-recognize images with adversarial training. Yeh et al. [54] used the chirality transform, a geometric transform that generates an antipode of a target, for pose regression. In this paper, we take a different approach based on mirror systems for unsupervised learning so that non-annotated images can be used to improve the performance.

2.2 Fidelity-Aware Image Processing

The mirror structure has been used successfully for various image processing tasks including domain conversion. A representative example is the VAE that jointly learns a generative model (decoder) of observed variables from latent variables following a prior distribution, and a recognition model (encoder) of the latent variables from the observed variables [24]. The VAE can generate new samples by randomly drawing latent variables from the prior distribution. CycleGAN [57], DiscoGAN [21], and DualGAN [55] are popular variants of GANs using mirror structures for image-to-image conversion. The key feature of these methods is to consider bidirectional inter-domain mappings from unpaired data. Qiao et al. [36] recently proposed MirrorGAN for bidirectional text-image conversion. Yildiri et al. [56] proposed an analysis-by-synthesis approach to joint 3D face generation and recognition from a cognitive point of view. The success of these methods indicates the potential of the mirror structure for stably training a DNN with non-annotated data.

In the context of pose estimation, we propose the first mirror-structured DNN for human pose estimation that integrates the two-level mirror systems in a hierarchically autoencoding manner. Recently, de Bem et al. [12] proposed a fidelity-aware pose estimation method based on an image-to-pose-to-image model called VAEGAN. This model can be trained in a semi-supervised manner, but its generalization capability has not been fully validated on a dataset with annotated and non-annotated images having various background images. The advantages of our method are that it can simultaneously consider both the pose plausibility and fidelity and that it can work robustly against changes in background images thanks to the foreground/background segmentation.

3. Proposed Method

This section describes the proposed method based on a fully probabilistic model of poses and images for 2D pose estimation in images of people (Fig. 2). MirrorNet is a hierarchical VAE that is one technique for amortized variational inference (AVI) [10], [24], [28], [38], and consists of a VAE of images (i.e., a *pose-to-image* generative model and an *image-to-pose* recognition model), and a VAE of poses (i.e., a *primitive-to-pose* generative model and a *pose-to-primitive* recognition model). In theory, this model can be trained in an unsupervised manner by using non-annotated images only, or by using unpaired images and poses. In practice, the model is trained in a semi-supervised

manner by using partially annotated images. Each model is first trained separately to stabilize the training, and then all models are jointly trained for further optimization. Once the training is completed, only the image-to-pose recognition model is used for pose estimation. The hierarchical autoencoding architecture is effective for estimating poses that are anatomically plausible and reproduce the original images with high fidelity.

3.1 Problem Specification

Let $\mathbf{X} = \{\mathbf{x}_n \in \mathbb{R}^{D^x}\}_{n=1}^N$ and $\mathbf{S} = \{\mathbf{s}_n \in \mathbb{R}^{D^s}\}_{n=1}^N$ be a set of images and a set of poses corresponding to \mathbf{X} , respectively, where D^x is the number of dimensions of each image, D^s is the number of dimensions of each pose, and N is the number of images. We assume that each \mathbf{x}_n is an RGB image featuring single or multiple persons, showing all or parts of their bodies, and each \mathbf{s}_n is a set of grayscale images, each of which represents the position of a joint using a heatmap [44]. Note that a coordinate taking the maximum pixel value in each of the grayscale images is retrieved as the 2D joint coordinate during a test phase.

Let $\mathbf{A} = \{\mathbf{a}_n \in \mathbb{R}^{D^a}\}_{n=1}^N$ and $\mathbf{G} = \{\mathbf{g}_n \in \mathbb{R}^{D^g}\}_{n=1}^N$ be a set of *appearances* representing the foreground features of \mathbf{X} (e.g., skin and hair colors and textures) and a set of *scenes* representing the background features of \mathbf{X} (e.g., places, color, and brightness), respectively, where D^a and D^g are the number of dimensions of the latent spaces. These latent features are used in combination with \mathbf{S} for representing \mathbf{X} . Let $\mathbf{Z} = \{\mathbf{z}_n \in \mathbb{R}^{D^z}\}_{n=1}^N$ be a set of *primitives* representing the features of \mathbf{S} (e.g., scales, positions, and orientations of joints), where D^z is the number of dimensions of the latent space.

Our goal is to train a pose recognizer that maps \mathbf{X} to \mathbf{S} . Let M be the number of annotated images. In a supervised condition, \mathbf{X} and \mathbf{S} are given as observed data ($M = N$). In an unsupervised condition, only \mathbf{X} is given ($M = 0$). In a semi-supervised condition, \mathbf{X} and a part of \mathbf{S} , i.e., $\{\mathbf{s}_n\}_{n=1}^M$, are given.

3.2 Generative Modeling

We formulate a unified hierarchical generative model of images \mathbf{X} , poses \mathbf{S} , appearances \mathbf{A} , scenes \mathbf{G} , and primitives \mathbf{Z} that integrates a deep generative model of \mathbf{X} from \mathbf{S} , \mathbf{A} , and \mathbf{G} with a deep generative model of \mathbf{S} from \mathbf{Z} as follows:

$$\begin{aligned} p(\mathbf{X}, \mathbf{S}, \mathbf{A}, \mathbf{G}, \mathbf{Z}) &= p_\theta(\mathbf{X}|\mathbf{S}, \mathbf{A}, \mathbf{G})p_\phi(\mathbf{S}|\mathbf{Z})p(\mathbf{A})p(\mathbf{G})p(\mathbf{Z}) \\ &= \prod_{n=1}^N p_\theta(\mathbf{x}_n|\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n)p_\phi(\mathbf{s}_n|\mathbf{z}_n)p(\mathbf{a}_n)p(\mathbf{g}_n)p(\mathbf{z}_n), \end{aligned} \quad (1)$$

where θ and ϕ are the sets of trainable parameters of the deep generative models of \mathbf{x}_n and \mathbf{s}_n , respectively. The pose likelihood $p_\theta(\mathbf{x}_n|\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n)$ evaluates the pose fidelity to the given images and the pose prior $p_\phi(\mathbf{s}_n|\mathbf{z}_n)$ prevents anatomically implausible pose estimates. The remaining terms are prior probability distributions of \mathbf{a}_n , \mathbf{g}_n , and \mathbf{z}_n .

The *pose-to-image* generation model $p_\theta(\mathbf{x}_n|\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n)$ and the *primitive-to-pose* generation model $p_\phi(\mathbf{s}_n|\mathbf{z}_n)$ are both formulated as follows:

$$p_\theta(\mathbf{x}_n|\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n) = \mathcal{N}(\mathbf{x}_n; \mu_\theta(\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n), \sigma_\theta^2(\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n)\mathbf{I}_{D^x}), \quad (2)$$

$$p_\phi(\mathbf{s}_n|\mathbf{z}_n) = \mathcal{N}(\mathbf{s}_n; \mu_\phi(\mathbf{z}_n), \sigma_\phi^2(\mathbf{z}_n)\mathbf{I}_{D^s}), \quad (3)$$

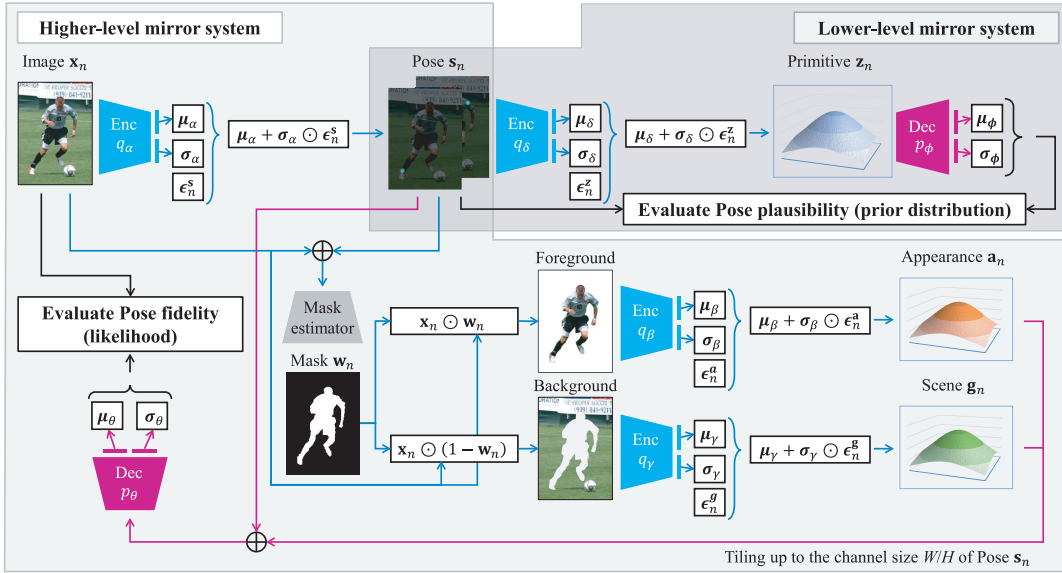


Fig. 2 The proposed architecture of MirrorNet integrating a VAE for poses (lower-level mirror system) with a pose-conditioned VAE for images (higher-level mirror system) in a hierarchical Bayesian manner. In terms of generative modeling, the decoder of the pose VAE serves as a prior distribution of poses $p(\mathbf{S})$ to evaluate the pose plausibility and the decoder of the image VAE as a likelihood function of poses $p(\mathbf{X}|\mathbf{S})$ to evaluate the pose fidelity. In terms of posterior inference, the encoder of the pose VAE is used as a variational posterior distribution of poses $q(\mathbf{S}|\mathbf{X})$. Such a statistical approach based on a complete probabilistic generative model enables *semi-supervised* pose estimation using any images with/without pose annotations.

where $\mu_\theta(\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n)$ and $\sigma_\theta(\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n)$ are the outputs of a DNN with parameter θ that takes \mathbf{s}_n , \mathbf{a}_n , and \mathbf{g}_n as input, and $\mu_\phi(\mathbf{z}_n)$ and $\sigma_\phi(\mathbf{z}_n)$ are the outputs of a DNN with parameters ϕ that takes \mathbf{z}_n as input. The priors $p(\mathbf{a}_n)$, $p(\mathbf{g}_n)$, and $p(\mathbf{z}_n)$ are set to the standard Gaussian distributions as follows:

$$p(\mathbf{a}_n) = \mathcal{N}(\mathbf{a}_n; \mathbf{0}_{D^a}, \mathbf{I}_{D^a}), \quad (4)$$

$$p(\mathbf{g}_n) = \mathcal{N}(\mathbf{g}_n; \mathbf{0}_{D^g}, \mathbf{I}_{D^g}), \quad (5)$$

$$p(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n; \mathbf{0}_{D^z}, \mathbf{I}_{D^z}), \quad (6)$$

where $\mathbf{0}_{D^\dagger}$ ($D^\dagger = \{D^a, D^g, D^z\}$) and \mathbf{I}_{D^\dagger} are the zero vector of size D^\dagger and the identity matrix of size D^\dagger , respectively.

3.3 Unsupervised Learning

We explain the unsupervised learning of the proposed model using only images \mathbf{X} , which is the basis for practical semi-supervised learning using *partially* annotated images (Section 3.4). Given a set of images \mathbf{X} as observed data, our goal is to infer the distribution of the latent variables $\mathbf{\Omega} \equiv (\mathbf{S}, \mathbf{A}, \mathbf{G}, \mathbf{Z})$. We estimate optimal parameters θ^* and ϕ^* in the framework of maximum likelihood estimation as follows:

$$\theta^*, \phi^* = \underset{\theta, \phi}{\operatorname{argmax}} p(\mathbf{X}), \quad (7)$$

and $p(\mathbf{X})$ is the marginal likelihood given by

$$p(\mathbf{X}) = \int p(\mathbf{X}, \mathbf{\Omega}) d\mathbf{\Omega}. \quad (8)$$

where $p(\mathbf{X}, \mathbf{\Omega})$ is the joint probability distribution given by Eq. (1).

Because Eq. (8) is analytically intractable, we use an amortized variational inference (AVI) technique [10], [24], [28], [38] that

introduces an arbitrary variational posterior distribution $q(\mathbf{\Omega}|\mathbf{X})$ and makes it approach as close as possible to the true posterior distribution $p(\mathbf{\Omega}|\mathbf{X})$ (Section 3.3.1). The minimization of the Kullback–Leibler (KL) divergence between these posteriors is equivalent to the maximization of a variational lower bound \mathcal{L} of $\log p(\mathbf{X})$ with respect to $q(\mathbf{\Omega}|\mathbf{X})$. Thus, the optimal parameters θ^* and ϕ^* can be obtained by maximizing the variational lower bound \mathcal{L} instead of $\log p(\mathbf{X})$ (Section 3.3.2).

3.3.1 Variational Lower Bound

Using Jensen’s inequality, a variational lower bound \mathcal{L}^X of $\log p(\mathbf{X})$ can be derived as follows:

$$\log p(\mathbf{X}) \geq \int q(\mathbf{\Omega}|\mathbf{X}) \log \frac{p(\mathbf{X}, \mathbf{\Omega})}{q(\mathbf{\Omega}|\mathbf{X})} d\mathbf{\Omega} \stackrel{\text{def}}{=} \mathcal{L}^X, \quad (9)$$

where the equality holds, i.e., \mathcal{L}^X is maximized, if and only if $q(\mathbf{\Omega}|\mathbf{X}) = p(\mathbf{\Omega}|\mathbf{X})$. Because this equality condition cannot be computed analytically, $q(\mathbf{\Omega}|\mathbf{X})$ is approximated by a factorized form as follows:

$$\begin{aligned} q(\mathbf{\Omega}|\mathbf{X}) &= q_\alpha(\mathbf{S}|\mathbf{X})q_\beta(\mathbf{A}|\mathbf{S}, \mathbf{X})q_\gamma(\mathbf{G}|\mathbf{S}, \mathbf{X})q_\delta(\mathbf{Z}|\mathbf{S}) \\ &= \prod_{n=1}^N q_\alpha(\mathbf{s}_n|\mathbf{x}_n)q_\beta(\mathbf{a}_n|\mathbf{s}_n, \mathbf{x}_n)q_\gamma(\mathbf{g}_n|\mathbf{s}_n, \mathbf{x}_n)q_\delta(\mathbf{z}_n|\mathbf{s}_n), \end{aligned} \quad (10)$$

where α , β , γ , and δ are the sets of parameters of these four variational distributions, respectively.

In the statistical framework of AVI, we introduce a DNN-based posterior distribution $q(\mathbf{\Omega}|\mathbf{X})$ such that the complex true posterior distribution $p(\mathbf{\Omega}|\mathbf{X})$ can be well approximated by $q(\mathbf{\Omega}|\mathbf{X})$. Specifically, we introduce a deep *image-to-pose* model $q_\alpha(\mathbf{s}_n|\mathbf{x}_n)$, a deep *image-to-appearance* model $q_\beta(\mathbf{a}_n|\mathbf{s}_n, \mathbf{x}_n)$, a deep *image-to-scene* model $q_\gamma(\mathbf{g}_n|\mathbf{s}_n, \mathbf{x}_n)$, and a deep *pose-to-primitive* model $q_\delta(\mathbf{z}_n|\mathbf{s}_n)$ as follows:

$$q_\alpha(\mathbf{s}_n|\mathbf{x}_n) = \mathcal{N}(\mathbf{s}_n; \mu_\alpha(\mathbf{x}_n), \sigma_\alpha^2(\mathbf{x}_n)\mathbf{I}_{D^s}), \quad (11)$$

$$q_\beta(\mathbf{a}_n|\mathbf{s}_n, \mathbf{x}_n) = \mathcal{N}(\mathbf{a}_n; \mu_\beta(\mathbf{s}_n, \mathbf{x}_n), \sigma_\beta^2(\mathbf{s}_n, \mathbf{x}_n)\mathbf{I}_{D^a}), \quad (12)$$

$$q_\gamma(\mathbf{g}_n|\mathbf{s}_n, \mathbf{x}_n) = \mathcal{N}(\mathbf{g}_n; \mu_\gamma(\mathbf{s}_n, \mathbf{x}_n), \sigma_\gamma^2(\mathbf{s}_n, \mathbf{x}_n)\mathbf{I}_{D^g}), \quad (13)$$

$$q_\delta(\mathbf{z}_n|\mathbf{s}_n) = \mathcal{N}(\mathbf{z}_n; \mu_\delta(\mathbf{s}_n), \sigma_\delta^2(\mathbf{s}_n)\mathbf{I}_{D^z}), \quad (14)$$

where $\mu_\alpha(\mathbf{x}_n)$ and $\sigma_\alpha(\mathbf{x}_n)$ are the outputs of a DNN with parameters α that takes \mathbf{x}_n as input, $\mu_{\ddagger}(\mathbf{s}_n, \mathbf{x}_n)$ and $\sigma_{\ddagger}(\mathbf{s}_n, \mathbf{x}_n)$ ($\ddagger = \beta$ or γ) are the outputs of a DNN with parameters \ddagger that takes \mathbf{s}_n and \mathbf{x}_n as input, and $\mu_\delta(\mathbf{s}_n)$ and $\sigma_\delta(\mathbf{s}_n)$ are the outputs of a DNN with parameters δ that takes \mathbf{s}_n as input.

Substituting both of the generative model given by Eq. (1) with Eqs. (2)–(6) and the recognition model given by Eq. (10) with Eqs. (11)–(14) into Eq. (9), the variational lower bound $\mathcal{L}^{\mathbf{X}}$ can be rewritten as the sum of $\{\mathcal{L}_n^{\mathbf{X}}\}_{n=1}^N$ ($\mathcal{L}^{\mathbf{X}} = \sum_n \mathcal{L}_n^{\mathbf{X}}$) as follows (Appendix A.1):

$$\begin{aligned} \mathcal{L}_n^{\mathbf{X}} &= \mathbb{E}_q[\log p_\theta(\mathbf{x}_n|\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n) + \log p_\phi(\mathbf{s}_n|\mathbf{z}_n) + \log p(\mathbf{a}_n) \\ &\quad + \log p(\mathbf{g}_n) + \log p(\mathbf{z}_n) - \log q_\alpha(\mathbf{s}_n|\mathbf{x}_n) \\ &\quad - \log q_\beta(\mathbf{a}_n|\mathbf{s}_n, \mathbf{x}_n) - \log q_\gamma(\mathbf{g}_n|\mathbf{s}_n, \mathbf{x}_n) - \log q_\delta(\mathbf{z}_n|\mathbf{s}_n)] \\ &= \mathbb{E}_q[\log p_\theta(\mathbf{x}_n|\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n)] + \mathbb{E}_q[\log p_\phi(\mathbf{s}_n|\mathbf{z}_n)] \\ &\quad - \mathbb{E}_q[\log q_\alpha(\mathbf{s}_n|\mathbf{x}_n)] - \mathbb{E}_q[\text{KL}(q_\beta(\mathbf{a}_n|\mathbf{s}_n, \mathbf{x}_n)||p(\mathbf{a}_n))] \\ &\quad - \mathbb{E}_q[\text{KL}(q_\gamma(\mathbf{g}_n|\mathbf{s}_n, \mathbf{x}_n)||p(\mathbf{g}_n))] - \mathbb{E}_q[\text{KL}(q_\delta(\mathbf{z}_n|\mathbf{s}_n)||p(\mathbf{z}_n))], \end{aligned} \quad (15)$$

where the first term represents the fidelity of a pose \mathbf{s}_n with an original image \mathbf{x}_n having features \mathbf{a}_n and \mathbf{g}_n , the second term represents the plausibility of \mathbf{s}_n , the third term prevents the overfitting of the recognition model α , and the fourth to sixth terms evaluate the similarities between the recognition models β , γ , and δ and the priors on \mathbf{a}_n , \mathbf{g}_n , and \mathbf{z}_n , respectively.

3.3.2 Parameter Optimization

Because Eq. (15) still includes intractable expectations, we perform Monte Carlo integration using samples \mathbf{s}_n , \mathbf{a}_n , \mathbf{g}_n , and \mathbf{z}_n obtained by *reparametrization trick* [24] as follows:

$$\epsilon_n^s \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{D^s}), \quad (16)$$

$$\epsilon_n^a \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{D^a}), \quad (17)$$

$$\epsilon_n^g \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{D^g}), \quad (18)$$

$$\epsilon_n^z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{D^z}), \quad (19)$$

$$\mathbf{s}_n = \mu_\alpha(\mathbf{x}_n) + \epsilon_n^s \odot \sigma_\alpha(\mathbf{x}_n), \quad (20)$$

$$\mathbf{a}_n = \mu_\beta(\mathbf{s}_n, \mathbf{x}_n) + \epsilon_n^a \odot \sigma_\beta(\mathbf{s}_n, \mathbf{x}_n), \quad (21)$$

$$\mathbf{g}_n = \mu_\gamma(\mathbf{s}_n, \mathbf{x}_n) + \epsilon_n^g \odot \sigma_\gamma(\mathbf{s}_n, \mathbf{x}_n), \quad (22)$$

$$\mathbf{z}_n = \mu_\delta(\mathbf{s}_n) + \epsilon_n^z \odot \sigma_\delta(\mathbf{s}_n), \quad (23)$$

where \odot indicates the element-wise product. Although in theory a sufficient number of samples should be generated to perform accurate Monte Carlo integration, we generate only one sample for each variable as in the standard VAE [24].

Using these techniques, the lower bound $\mathcal{L}^{\mathbf{X}}$ given by Eq. (15) can be approximately computed, and can thus be maximized with respect to θ , ϕ , α , β , γ , and δ (Fig. 2). First, the recognition models α , β , γ , and δ are used to *deterministically* generate samples \mathbf{s}_n , \mathbf{a}_n , \mathbf{g}_n , and \mathbf{z}_n in Eqs. (20)–(23), and to calculate the last four regularization terms of Eq. (15), respectively. Given the samples \mathbf{s}_n , \mathbf{a}_n , \mathbf{g}_n , and \mathbf{z}_n , the generative models θ and ϕ are used to

calculate the first two reconstruction terms of Eq. (15), respectively. The recognition models α , β , γ , and δ , and the generative models ϕ and θ can thus be concatenated in this order with the reparametrization trick given by Eqs. (20)–(23), and are jointly optimized in an autoencoding manner with an objective function given by Eq. (15).

3.4 Supervised Learning

We explain the supervised learning of the proposed model using paired data of \mathbf{X} and \mathbf{S} . This approach follows the manner of the semi-supervised learning of a VAE [23]. While the variational lower bound $\mathcal{L}^{\mathbf{X}}$ of $\log p(\mathbf{X})$ is maximized in the unsupervised condition (Section 3.3), we aim to maximize the variational lower bound $\mathcal{L}^{\mathbf{X}, \mathbf{S}}$ of $\log p(\mathbf{X}, \mathbf{S})$, which is given by

$$\log p(\mathbf{X}, \mathbf{S}) \geq \int q(\Theta|\mathbf{S}, \mathbf{X}) \log \frac{p(\mathbf{X}, \Theta)}{q(\Theta|\mathbf{S}, \mathbf{X})} d\Theta \stackrel{\text{def}}{=} \mathcal{L}^{\mathbf{X}, \mathbf{S}}, \quad (24)$$

where $\Theta = \Omega|\mathbf{S} = \{\mathbf{A}, \mathbf{G}, \mathbf{Z}\}$. As in Eq. (10), $q(\Theta|\mathbf{S}, \mathbf{X})$ is factorized as

$$\begin{aligned} q(\Theta|\mathbf{S}, \mathbf{X}) &= q_\beta(\mathbf{A}|\mathbf{S}, \mathbf{X})q_\gamma(\mathbf{G}|\mathbf{S}, \mathbf{X})q_\delta(\mathbf{Z}|\mathbf{S}) \\ &= \prod_{n=1}^N q_\beta(\mathbf{a}_n|\mathbf{s}_n, \mathbf{x}_n)q_\gamma(\mathbf{g}_n|\mathbf{s}_n, \mathbf{x}_n)q_\delta(\mathbf{z}_n|\mathbf{s}_n), \end{aligned} \quad (25)$$

where $q_\beta(\mathbf{a}_n|\mathbf{s}_n, \mathbf{x}_n)$, $q_\gamma(\mathbf{g}_n|\mathbf{s}_n, \mathbf{x}_n)$, and $q_\delta(\mathbf{z}_n|\mathbf{s}_n)$ are given by Eqs. (12)–(14), respectively. Substituting both of the probabilistic model given by Eq. (1) with Eqs. (2)–(6) and the inference model given by Eq. (25) with Eqs. (12)–(14) into Eq. (24), $\mathcal{L}^{\mathbf{X}, \mathbf{S}}$ can be rewritten as the sum of $\{\mathcal{L}_n^{\mathbf{X}, \mathbf{S}}\}_{n=1}^N$ ($\mathcal{L}^{\mathbf{X}, \mathbf{S}} = \sum_n \mathcal{L}_n^{\mathbf{X}, \mathbf{S}}$) as follows (Appendix A.1):

$$\begin{aligned} \mathcal{L}_n^{\mathbf{X}, \mathbf{S}} &= \mathbb{E}_q[\log p_\theta(\mathbf{x}_n|\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n) + \log p_\phi(\mathbf{s}_n|\mathbf{z}_n) + \log p(\mathbf{a}_n) \\ &\quad + \log p(\mathbf{g}_n) + \log p(\mathbf{z}_n) - \log q_\beta(\mathbf{a}_n|\mathbf{s}_n, \mathbf{x}_n) \\ &\quad - \log q_\gamma(\mathbf{g}_n|\mathbf{s}_n, \mathbf{x}_n) - \log q_\delta(\mathbf{z}_n|\mathbf{s}_n)] \\ &= \mathbb{E}_q[\log p_\theta(\mathbf{x}_n|\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n)] + \mathbb{E}_q[\log p_\phi(\mathbf{s}_n|\mathbf{z}_n)] \\ &\quad - \text{KL}(q_\beta(\mathbf{a}_n|\mathbf{s}_n, \mathbf{x}_n)||p(\mathbf{a}_n)) - \text{KL}(q_\gamma(\mathbf{g}_n|\mathbf{s}_n, \mathbf{x}_n)||p(\mathbf{g}_n)) \\ &\quad - \text{KL}(q_\delta(\mathbf{z}_n|\mathbf{s}_n)||p(\mathbf{z}_n)). \end{aligned} \quad (26)$$

A major problem in such supervised learning is that the recognition model $q_\alpha(\mathbf{s}_n|\mathbf{x}_n)$, which plays a central role in human pose estimation from images, cannot be trained because it does not appear in Eq. (26). To solve this problem, we add a term to assess the predictive performance of $q_\alpha(\mathbf{s}_n|\mathbf{x}_n)$ to \mathcal{L}_n , following [23] as

$$\mathcal{L}_{n, \lambda}^{\mathbf{X}, \mathbf{S}} = \mathcal{L}_n^{\mathbf{X}, \mathbf{S}} + \lambda \log q_\alpha(\mathbf{s}_n|\mathbf{x}_n), \quad (27)$$

$$\begin{aligned} &\log q_\alpha(\mathbf{s}_n|\mathbf{x}_n) \\ &= -\frac{1}{2} \sum_{d_s=1}^{D^s} \left(\log(2\pi\sigma_{\alpha, d_s}^2(\mathbf{x}_n)) + \frac{(\mathbf{s}_n - \mu_{\alpha, d_s}(\mathbf{x}_n))^2}{\sigma_{\alpha, d_s}^2(\mathbf{x}_n)} \right), \end{aligned} \quad (28)$$

where λ is a hyperparameter that controls the balance between purely generative learning and purely discriminative learning. In our method, we empirically used $\lambda = 0.01$ in all experiments. The new objective function $\mathcal{L}_\lambda^{\mathbf{X}, \mathbf{S}} = \sum_n \mathcal{L}_{n, \lambda}^{\mathbf{X}, \mathbf{S}}$ can be maximized with respect to θ , ϕ , α , β , γ , and δ jointly in the same way as the unsupervised learning described in Section 3.3.2, where \mathbf{a}_n , \mathbf{g}_n , and \mathbf{z}_n are obtained by using Eqs. (21)–(23), and \mathbf{s}_n is given.

3.5 Semi-supervised Learning

In the semi-supervised condition, where \mathbf{X} is only partially annotated, we define a new objective function \mathcal{L} by accumulating $\mathcal{L}_n^{\mathbf{X}}$ used for unsupervised learning or $\mathcal{L}_{n,\lambda}^{\mathbf{X},\mathbf{S}}$ used for supervised learning as follows:

$$\mathcal{L} \stackrel{\text{def}}{=} \sum_{n: \mathbf{x}_n \text{ is given}} \mathcal{L}_n^{\mathbf{X}} + \eta \sum_{n: \mathbf{x}_n \text{ \& } \mathbf{s}_n \text{ are given}} \mathcal{L}_{n,\lambda}^{\mathbf{X},\mathbf{S}}, \quad (29)$$

where η is a weighting factor, which was set to 1 in our experiments reported in Section 5, because this choice is theoretically reasonable in terms of probabilistic modeling and we found no significant difference in pose estimation performance between $\eta = 0.25, 0.50, 1.0, 2.0, 4.0$. All generation and recognition models can be trained for all samples regardless of the availability of their annotations. We will discuss the effectiveness of the curriculum learning based on the pre-training of each component of MirrorNet and the joint training of the whole MirrorNet in Section 5.4.

4. Implementation

This section describes the implementation of MirrorNet, which is based on curriculum learning. First, we separately pre-train the components of MirrorNet, i.e., the pose recognizer α (Section 4.1), the pose-conditioned image VAE with the generator θ and the recognizers β and γ (Section 4.2), and the pose VAE with the generator ϕ and the recognizer δ (Section 4.3). We then train the whole MirrorNet under a supervised condition (Section 3.4) and further optimize it under a semi-supervised condition (Section 3.5).

Note that, as shown in Fig. 2, the pose-conditioned image VAE has a human mask estimator as a subcomponent for separating an image into foreground and background images which helps to stabilize the training of MirrorNet.

4.1 Pose Recognizer

The image-to-pose recognizer α is the most fundamental part of pose estimation, and is pre-trained in a *supervised* manner by using paired data of \mathbf{X} and \mathbf{S} . We maximize an objective function given by

$$\mathcal{L}_n(\alpha) \stackrel{\text{def}}{=} \log q_\alpha(\mathbf{s}_n | \mathbf{x}_n), \quad (30)$$

where the variance $\sigma_\alpha^2(\mathbf{x}_n)$ is fixed to 0.01 for stability.

The network α can be implemented with any DNN that outputs the heatmaps of joint positions, e.g., a stack of eight residual hourglass networks (HGNet) [30], ResNet-50 [50], and high-resolution sub-networks (HRNet) [42].

4.2 Pose-Conditioned Image VAE

The pose-conditioned image VAE consisting of the image-to-appearance recognizer β , the image-to-scene recognizer γ (encoders), and the appearance/scene-to-image generator θ (decoder) is pre-trained in an *unsupervised* manner by using paired data of \mathbf{X} and \mathbf{S} . We maximize a variational lower bound $\mathcal{L}(\theta, \beta, \gamma)$ of the marginal log likelihood $\log p(\mathbf{X} | \mathbf{S})$. More specifically, we have

$$\log p(\mathbf{x}_n | \mathbf{s}_n) \geq \mathbb{E}_q [\log p_\theta(\mathbf{x}_n | \mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n) + \log p(\mathbf{a}_n) + \log p(\mathbf{g}_n)]$$

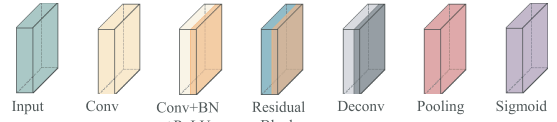


Fig. 3 Layers used for implementing DNNs.

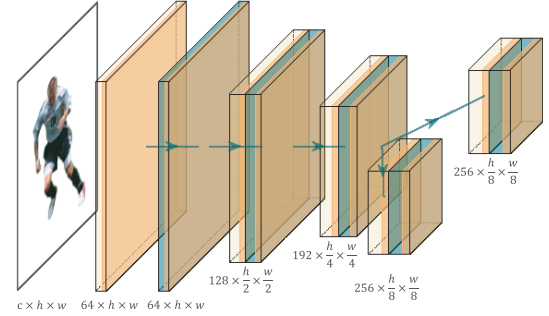


Fig. 4 The network configuration common in the recognizers β and γ of the pose-conditioned image VAE taking as input foreground and background images \mathbf{x}_n^{fg} and \mathbf{x}_n^{bg} ($c = 3$), respectively, and the recognizer δ of the pose VAE taking as input the pose \mathbf{s}_n (c is the number of joints).

$$\begin{aligned} & -\log q_\beta(\mathbf{a}_n | \mathbf{s}_n, \mathbf{x}_n) - \log q_\gamma(\mathbf{g}_n | \mathbf{s}_n, \mathbf{x}_n) \\ & = \mathbb{E}_q [\log p_\theta(\mathbf{x}_n | \mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n)] - \text{KL}(q_\beta(\mathbf{a}_n | \mathbf{s}_n, \mathbf{x}_n) \| p(\mathbf{a}_n)) \\ & \quad - \text{KL}(q_\gamma(\mathbf{g}_n | \mathbf{s}_n, \mathbf{x}_n) \| p(\mathbf{g}_n)) \\ & \stackrel{\text{def}}{=} \mathcal{L}_n(\theta, \beta, \gamma), \end{aligned} \quad (31)$$

where $\mathcal{L}(\theta, \beta, \gamma) = \sum_{n=1}^N \mathcal{L}_n(\theta, \beta, \gamma)$. The three networks θ , β , and γ can be optimized jointly by using the reparametrization tricks [24] given by Eq. (21) and Eq. (22), where the variance $\sigma_\theta^2(\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n)$ of the generator θ is fixed to 1 for stability.

To encourage the disentanglement between the foreground features (appearance) \mathbf{a}_n and the background features (scene) \mathbf{g}_n , we separately input foreground and background parts of the original image \mathbf{x}_n into the two encoders β and γ , respectively, instead of directly feeding \mathbf{x}_n into β and γ . Specifically, an image $\mathbf{x}_n^* \in \mathbb{R}^{D^*}$, a reduced-size version of \mathbf{x}_n , is first split into foreground and background images \mathbf{x}_n^{fg} and \mathbf{x}_n^{bg} as follows:

$$\mathbf{x}_n^{\text{fg}} = \mathbf{x}_n^* \odot \mathbf{w}_n, \quad (32)$$

$$\mathbf{x}_n^{\text{bg}} = \mathbf{x}_n^* \odot (\mathbf{1} - \mathbf{w}_n), \quad (33)$$

where \odot indicates the element-wise product and $\mathbf{w}_n \in \mathbb{R}^{D^*}$ represents a mask image estimated from \mathbf{x}_n^* with the additional information of the pose \mathbf{s}_n . In this paper, we use a neural mask estimator ψ trained in a supervised manner such that the mean squared error between the estimated and ground-truth masks is minimized.

The recognizers β and γ are implemented as stacks of four residual blocks [16] (Fig. 3 and Fig. 4). Unlike the original ResNet, a branching architecture is introduced in the last layer to output the mean and variance of the posterior Gaussian distribution. The generator θ is implemented with a U-Net [39] that takes as input a stack of the heatmaps of the joints given by \mathbf{s}_n and the latent variables \mathbf{a}_n and \mathbf{g}_n , where a branching architecture is introduced in the last layer to evaluate the pose fidelity with \mathbf{x}_n (Fig. 5). The mask estimator ψ is implemented as a U-Net [39] that takes as input a shrunk image \mathbf{x}_n^* and a stack of

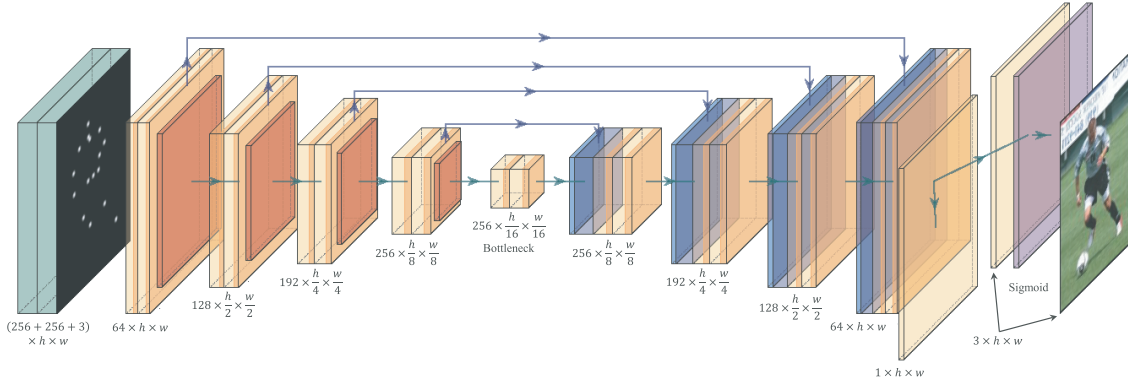


Fig. 5 The network configuration of the generator θ of the pose-conditioned image VAE taking as input the pose \mathbf{s}_n , the appearance \mathbf{a}_n , and the scene \mathbf{g}_n and yielding the mean $\mu_\theta(\mathbf{a}_n, \mathbf{g}_n, \mathbf{s}_n)$ and the variance $\sigma_\theta(\mathbf{a}_n, \mathbf{g}_n, \mathbf{s}_n)$.

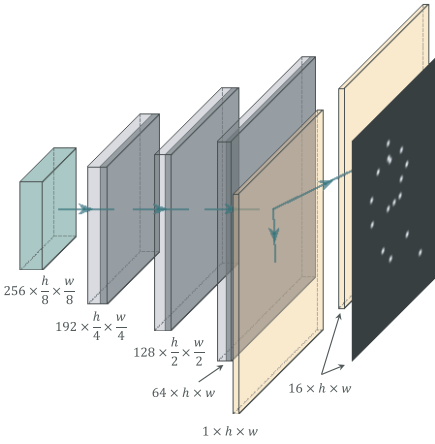


Fig. 6 The network configuration of the generator θ of the pose VAE taking as input the primitives \mathbf{z}_n and yielding the mean $\mu_\theta(\mathbf{z}_n)$ and variance $\sigma_\theta(\mathbf{z}_n)$.

the heatmaps of the joints given by \mathbf{s}_n and outputs a mask image \mathbf{w}_n . To obtain sharper mask images, we apply a sigmoid function, $\zeta(x) = (1 + \exp(-10x))^{-1}$, to every element of the output \mathbf{w}_n of the pre-trained estimator ψ .

4.3 Pose VAE

The pose VAE consisting of the pose-to-primitive recognizer δ and the primitive-to-pose generator ϕ (decoder) is pre-trained in an *unsupervised* manner by using only \mathbf{S} . We maximize a variational lower bound $\mathcal{L}(\phi, \delta)$ of the marginal log likelihood $\log p(\mathbf{S})$ evaluating the pose plausibility. More specifically, we have

$$\begin{aligned} \log p(\mathbf{s}_n) &\geq \mathbb{E}_q[\log p_\phi(\mathbf{s}_n|\mathbf{z}_n) + \log p(\mathbf{z}_n) - \log q_\delta(\mathbf{z}_n|\mathbf{s}_n)] \\ &= \mathbb{E}_q[\log p_\phi(\mathbf{s}_n|\mathbf{z}_n)] - \text{KL}(q_\delta(\mathbf{z}_n|\mathbf{s}_n)||p(\mathbf{z}_n)) \\ &\stackrel{\text{def}}{=} \mathcal{L}_n(\phi, \delta), \end{aligned} \quad (34)$$

where $\mathcal{L}(\phi, \delta) = \sum_{n=1}^N \mathcal{L}_n(\phi, \delta)$. The two networks ϕ and δ are optimized jointly with the reparametrization trick [24] (Eq. (23)), where the variance σ_ϕ^2 of the generator ϕ is fixed to 1 for stability.

The recognizer δ is implemented in the same way as the recognizers β and γ except for the input dimension (Fig. 4). The generator ϕ is implemented as a three-layered transposed convolutional network, where a branching architecture was introduced in the last layer to evaluate the pose plausibility (Fig. 6).

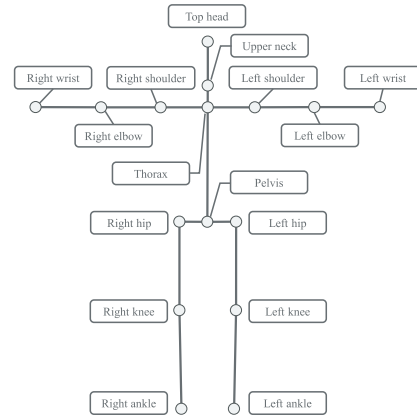


Fig. 7 Sixteen joints dealt with in our experiments.

5. Evaluation

This section reports comparative experiments conducted for evaluating the effectiveness of our semi-supervised plausibility- and fidelity-aware pose estimation method. Our goal is to train a DNN-based pose recognizer α that detects the coordinates of 16 joints (left and right ankles, knees, and hips, left and right wrists, elbows, and shoulders, pelvis, thorax, upper neck, and head top as shown in Fig. 7) from an image. We here validate two hypotheses: (A) under a *supervised* condition, the joint training of the recognition and generative models $\alpha, \beta, \gamma, \theta$, and ϕ outperforms the standalone training of the pose recognizer α , and (B) under a *semi-supervised* condition, the mirror architecture makes effective use of non-annotated images for improving performance.

5.1 Datasets and Criteria

We used two standard datasets that have been widely used in conventional studies on pose estimation.

5.1.1 Leeds Sports Pose (LSP) Dataset

The LSP dataset with its extension [18], [19] contains 12 K images of sports activities (11 K for training and 1 K for testing) in total. Each image originally has an annotation about the coordinates of the 14 joints except for the pelvis and thorax. In order to use the same set of joints as the MPII Human Pose Dataset [1] (see Section 5.1.2) and use the same configuration of MirrorNet, we estimated the ground-truth coordinate of the pelvis by aver-

aging the coordinates of the left and right hips. Similarly, we estimated the ground-truth coordinate of the thorax by averaging the coordinates of the left and right shoulders. Each image was cropped to a square region centering on a person and then scaled to $D^x = 256 \times 256$.

The performance of pose estimation was measured with the *percentage of correct keypoints* (PCK) [53]. The estimated coordinate of a joint was judged as correct if within τl_t pixels from the ground-truth coordinate, where l_t is the torso size defined as the diagonal length of the ground-truth bounding box of the torso and τ is a relative error tolerance ($\tau = 0.2$ in our experiment).

5.1.2 MPII Human Pose (MPII) Dataset

The MPII dataset [1] contains around 25 K images of daily activities (22 K for training and 3 K for testing). Each image annotated with the coordinates of the 16 joints was cropped to a square region centering on a person, and then scaled to $D^x = 256 \times 192$.

The performance of pose estimation was measured with the *percentage of correct keypoints in relation to head segment length* (PCKh) [1]. The estimated coordinate of each joint was judged as correct if within τl_h pixels around the ground-truth coordinate, where τ is a constant threshold and l_h is the head size corresponding to 60% of the diagonal length of the ground-truth head bounding box. We used $\tau = 0.5$ in our experiment.

5.2 Training Procedures

We randomly selected 20%, 40%, 60%, 80%, and 100% of the training data as annotated images and regarded the remaining part as non-annotated images. Only the annotated images were used for supervised training and the entire training data were used for semi-supervised training. As in the official implementation of Ref. [42], the training data were augmented with random scaling, rotation, and horizontal flipping [49]. The target data of s_n were made by stacking 16 reduced-size grayscale images (heatmaps) indicating the coordinates of the 16 joints ($D^s = 16 \times 64 \times 64$ or $16 \times 64 \times 48$). In the test phase, a coordinate taking the maximum value in each of the 16 grayscale images was detected.

We conducted curriculum learning as described in Section 4 and shown in Fig. 8, where the dimensions of the latent foreground, background and pose features were set to $D^a = D^b = D^z = 256 \times 8 \times 8$ or $256 \times 8 \times 6$ (Fig. 8).

(1) **Supervised pre-training:** The six sub-networks were trained independently in a *supervised* manner using the annotated images. The pose recognizer α based on the residual hourglass network (HGNet) [30], ResNet-50 [50], or the high-resolution network (HRNet) [42] was trained for 100 epochs (Section 4.1). The pose-conditioned image VAE consisting of the generator θ and the recognizers β and γ was trained for 200 epochs (Section 4.2). The pose VAE consisting of the generator ϕ and the recognizer δ was also trained for 200 epochs (Section 4.3). The mask estimator ψ was trained by using the UPI-S1h dataset [25] containing human images with silhouette annotations (i.e., masks), where images included in the LSP or MPII datasets were excluded.

(a) MirrorNet was built by combining the six sub-networks and the mask estimator ψ and passed to the step (2).

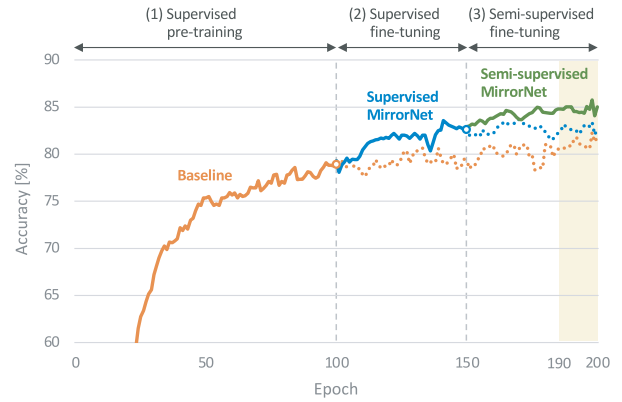


Fig. 8 Learning curves obtained with the pose recognizer α based on the hourglass network [30] on the LSP dataset, in which 40% of the training data were regarded as annotated images.

(b) The pose recognizer α was further trained for 100 epochs (i.e., 200 epochs in total) and the parameters at the last 10 epochs were used for evaluation (**baseline**).

(2) **Supervised fine-tuning:** MirrorNet was trained in a *supervised* manner with the same annotated images for 50 epochs, where the mask estimator ψ was not updated.

(a) MirrorNet at the last epoch was passed to the step (3).

(b) MirrorNet was further trained for 50 epochs and the parameters of the pose recognizer α at the last 10 epochs were used for evaluation (**supervised MirrorNet**).

(3) **Semi-supervised fine-tuning:** MirrorNet was further trained in a *semi-supervised* manner with the annotated and non-annotated images for 50 epochs, where the mask estimator ψ was not updated. The parameters at the last 10 epochs were used for evaluation (**semi-supervised MirrorNet**).

To make a fair comparison, the pose recognizer α was trained for 200 epochs in total under any conditions. The performance of pose estimation was measured by averaging the values of PCK@0.2 or PCKh@0.5 over the last 10 epochs.

All networks were implemented using PyTorch [33] and optimized using Adam [22] with a learning rate of 0.001. The mini-batch size was set to 128 images, which are fully annotated in the supervised training phase or consisted of 96 annotated images and 32 non-annotated images in the semi-supervised training phase.

5.3 Experimental Results

Tables 1, 2, 3, 4, 5 and 6 show the respective performance of pose estimation obtained by the pose recognizer α (Refs. [30], [50], or Ref. [42]) trained in the three ways (baseline, supervised MirrorNet, and semi-supervised MirrorNet) on the LSP and MPII datasets, respectively, and Fig. 9 comparatively shows the respective performance listed in the “Total” columns of Tables 1–6. Under all conditions, the supervised MirrorNet outperformed the baseline method by 4.61 ± 1.70 points on the LSP dataset and 4.91 ± 1.75 points on the MPII dataset, where the means and standard deviations were computed over the fifteen conditions, i.e., all possible combinations of the pose recognizers [30], [42], [50] and the five ratios of annotated images (20%, 40%, 60%, 80%, and 100%). The left five columns of Fig. 9 clearly show that the supervised MirrorNet significantly outperformed the baseline

Table 1 Pose estimation performance on the LSP dataset [18], [19] with the pose recognizer α based on HGNet [30].

	Training data		PCK@0.2							
	#annotated	#non-annotated	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Baseline [30]	2200	-	92.87	81.50	73.05	70.08	71.81	67.73	65.85	74.94
Supervised MirrorNet	2200	-	93.81	83.83	75.78	73.26	75.18	72.41	68.91	77.80
Semi-supervised MirrorNet	2200	8800	90.68	79.49	69.60	67.97	69.15	66.69	63.33	72.66
Baseline [30]	4400	-	92.51	85.46	79.75	77.67	76.60	77.91	74.49	80.81
Supervised MirrorNet	4400	-	95.07	86.55	80.02	76.08	80.36	79.71	75.64	82.10
Semi-supervised MirrorNet	4400	6600	94.88	88.94	83.32	80.27	82.34	82.41	79.15	84.68
Baseline [30]	6600	-	94.79	87.07	80.59	77.82	80.46	79.36	75.73	82.46
Supervised MirrorNet	6600	-	95.27	89.94	85.90	83.66	83.53	84.82	82.37	86.69
Semi-supervised MirrorNet	6600	4400	95.60	89.39	85.91	83.60	84.06	85.13	83.13	86.87
Baseline [30]	8800	-	94.73	87.83	81.88	79.24	82.06	82.35	79.05	84.10
Supervised MirrorNet	8800	-	95.62	89.71	85.10	83.84	84.51	86.29	83.68	87.19
Semi-supervised MirrorNet	8800	2200	95.63	89.98	85.65	83.98	84.87	86.29	83.44	87.34
Baseline [30]	11000	-	95.85	89.68	84.89	83.25	85.39	86.24	82.41	86.97
Supervised MirrorNet	11000	-	96.79	92.31	89.15	87.64	89.94	90.29	88.30	90.76

Table 2 Pose estimation performance on the LSP dataset [18], [19] with the pose recognizer α based on ResNet [50].

	Training data		PCK@0.2							
	#annotated	#non-annotated	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Baseline [50]	2200	-	86.69	69.26	56.62	52.32	57.22	53.37	48.04	60.83
Supervised MirrorNet	2200	-	89.25	76.55	65.97	61.14	65.67	62.69	54.02	68.19
Semi-supervised MirrorNet	2200	8800	85.15	75.49	63.86	58.69	63.58	57.17	47.66	64.89
Baseline [50]	4400	-	89.59	78.16	68.97	63.28	67.74	63.31	55.02	69.68
Supervised MirrorNet	4400	-	90.82	81.23	72.19	67.94	71.68	67.53	60.05	73.30
Semi-supervised MirrorNet	4400	6600	88.39	79.58	70.78	66.42	68.87	64.67	59.78	71.54
Baseline [50]	6600	-	89.59	78.40	69.26	64.51	70.32	64.47	57.23	70.82
Supervised MirrorNet	6600	-	92.50	83.15	75.29	72.23	74.39	71.70	63.62	76.35
Semi-supervised MirrorNet	6600	4400	92.36	83.38	75.31	72.24	75.55	73.27	64.85	77.00
Baseline [50]	8800	-	90.29	78.70	68.40	64.70	69.70	64.79	57.62	70.90
Supervised MirrorNet	8800	-	93.06	84.04	76.79	73.48	76.81	75.11	65.80	78.14
Semi-supervised MirrorNet	8800	2200	93.98	85.24	76.51	72.78	77.08	74.96	66.84	78.40
Baseline [50]	11000	-	92.35	82.93	75.11	69.71	78.66	72.57	62.79	76.60
Supervised MirrorNet	11000	-	94.33	86.94	80.91	76.98	84.64	80.00	72.42	82.56

Table 3 Pose estimation performance on the LSP dataset [18], [19] with the pose recognizer α based on HRNet [42].

	Training data		PCK@0.2							
	#annotated	#non-annotated	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Baseline [42]	2200	-	92.18	81.46	73.15	70.59	72.84	71.28	68.06	75.86
Supervised MirrorNet	2200	-	94.04	83.92	78.19	76.61	75.90	74.82	72.30	79.61
Semi-supervised MirrorNet	2200	8800	92.55	84.16	77.90	76.58	73.57	76.14	74.06	79.49
Baseline [42]	4400	-	93.05	83.40	76.19	74.31	75.77	73.53	70.78	78.36
Supervised MirrorNet	4400	-	95.02	88.23	82.45	81.23	81.30	81.16	78.76	84.22
Semi-supervised MirrorNet	4400	6600	94.47	88.48	83.31	82.57	81.63	82.95	80.23	85.02
Baseline [42]	6600	-	93.86	84.90	77.74	75.29	78.69	77.49	74.78	80.65
Supervised MirrorNet	6600	-	95.65	88.82	84.34	82.74	83.42	84.74	82.36	86.22
Semi-supervised MirrorNet	6600	4400	95.69	89.40	85.11	83.37	83.28	84.58	83.23	86.58
Baseline [42]	8800	-	93.99	85.49	79.21	77.20	77.96	78.57	75.70	81.43
Supervised MirrorNet	8800	-	95.61	89.69	85.21	83.90	83.98	85.31	83.67	86.98
Semi-supervised MirrorNet	8800	2200	95.70	89.91	85.52	83.60	82.96	85.78	83.81	86.95
Baseline [42]	11000	-	95.74	90.45	86.20	83.08	88.74	86.92	83.80	88.06
Supervised MirrorNet	11000	-	96.65	92.78	89.71	87.44	91.72	91.57	89.90	91.58

Table 4 Pose estimation performance on the MPII dataset[1] with the pose recognizer α based on HGNet [30].

	Training data		PCKh@0.5							
	#annotated	#non-annotated	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Baseline [30]	4449	-	93.26	86.86	76.10	69.17	75.07	67.82	63.91	76.91
Supervised MirrorNet	4449	-	94.10	89.27	79.05	72.89	77.64	71.16	67.60	79.63
Semi-supervised MirrorNet	4449	17797	93.34	87.92	77.00	70.03	73.60	67.68	63.02	77.02
Baseline [30]	8899	-	94.32	89.07	78.58	71.36	78.48	71.29	67.29	79.43
Supervised MirrorNet	8899	-	95.12	91.40	81.97	75.57	81.98	75.14	71.20	82.51
Semi-supervised MirrorNet	8899	13347	95.19	92.15	83.27	77.03	81.57	76.30	72.82	83.32
Baseline [30]	13347	-	94.93	91.24	81.63	74.97	81.92	74.41	69.83	82.07
Supervised MirrorNet	13347	-	95.92	93.37	84.96	78.54	85.07	78.80	74.90	85.18
Semi-supervised MirrorNet	13347	8899	95.85	93.56	85.06	79.15	85.35	79.26	75.47	85.47
Baseline [30]	17797	-	94.90	91.32	81.62	74.70	81.06	74.39	70.32	81.94
Supervised MirrorNet	17797	-	95.85	93.45	85.40	79.32	85.24	79.32	75.69	85.54
Semi-supervised MirrorNet	17797	4449	95.76	93.36	85.14	79.21	84.79	79.29	75.64	85.38
Baseline [30]	22246	-	95.08	91.94	82.47	75.98	82.65	75.69	71.58	82.94
Supervised MirrorNet	22246	-	96.13	94.01	86.47	80.05	86.41	80.56	76.85	86.40

Table 5 Pose estimation performance on the MPII dataset[1] with the pose recognizer α based on ResNet [50].

	Training data		PCKh@0.5							
	#annotated	#non-annotated	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Baseline [50]	4449	-	88.71	80.11	64.74	54.85	66.30	54.39	52.14	67.10
Supervised MirrorNet	4449	-	90.71	82.99	69.26	60.25	70.08	59.96	56.31	71.03
Semi-supervised MirrorNet	4449	17797	89.74	82.29	67.98	58.50	66.54	57.14	53.70	69.16
Baseline [50]	8899	-	90.56	82.91	68.39	58.58	70.61	59.16	55.31	70.50
Supervised MirrorNet	8899	-	92.68	87.16	74.86	65.74	76.39	66.54	61.74	76.01
Semi-supervised MirrorNet	8899	13347	92.94	87.31	74.76	66.50	75.14	66.18	61.26	75.87
Baseline [50]	13347	-	90.64	83.58	68.52	58.19	71.63	59.38	55.62	70.79
Supervised MirrorNet	13347	-	92.96	87.84	74.80	65.40	77.32	66.72	62.22	76.32
Semi-supervised MirrorNet	13347	8899	93.02	88.14	75.28	66.15	77.89	67.53	62.67	76.79
Baseline [50]	17797	-	89.49	81.37	65.70	54.90	68.78	56.91	53.50	68.40
Supervised MirrorNet	17797	-	92.36	86.68	73.50	64.19	76.36	65.63	60.34	75.19
Semi-supervised MirrorNet	17797	4449	92.72	87.58	74.04	64.20	76.88	65.92	60.65	75.61
Baseline [50]	22246	-	91.07	84.86	70.31	60.31	72.61	60.84	56.72	72.11
Supervised MirrorNet	22246	-	93.76	89.36	76.86	67.90	79.12	68.87	63.65	78.06

Table 6 Pose estimation performance on the MPII dataset [1] with the pose recognizer α based on HR-Net [42].

	Training data		PCKh@0.5							
	#annotated	#non-annotated	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Baseline [42]	4449	-	92.89	86.95	75.44	68.99	74.50	66.80	62.73	76.42
Supervised MirrorNet	4449	-	93.89	89.40	79.90	73.70	77.76	72.07	67.74	80.04
Semi-supervised MirrorNet	4449	17797	93.92	88.89	79.05	72.40	75.14	70.59	66.22	78.91
Baseline [42]	8899	-	93.86	88.27	77.52	70.69	77.62	69.71	66.05	78.52
Supervised MirrorNet	8899	-	95.32	91.82	83.09	76.67	82.71	76.44	72.43	83.35
Semi-supervised MirrorNet	8899	13347	95.32	91.97	83.09	76.77	81.31	76.08	72.26	83.12
Baseline [42]	13347	-	93.83	88.23	78.02	70.78	77.87	69.92	66.26	78.70
Supervised MirrorNet	13347	-	95.84	92.67	84.60	78.04	84.03	77.62	73.76	84.48
Semi-supervised MirrorNet	13347	8899	95.67	92.59	84.67	78.21	84.15	77.95	73.90	84.57
Baseline [42]	17797	-	93.51	87.55	76.84	69.50	75.97	67.84	63.48	77.31
Supervised MirrorNet	17797	-	95.71	92.93	84.31	77.73	83.89	77.41	73.16	84.30
Semi-supervised MirrorNet	17797	4449	95.60	92.81	84.31	77.55	84.09	77.50	73.26	84.30
Baseline [42]	22246	-	92.55	87.22	75.95	69.69	75.68	67.11	63.67	76.92
Supervised MirrorNet	22246	-	95.91	93.77	85.85	79.33	85.48	79.41	75.35	85.69

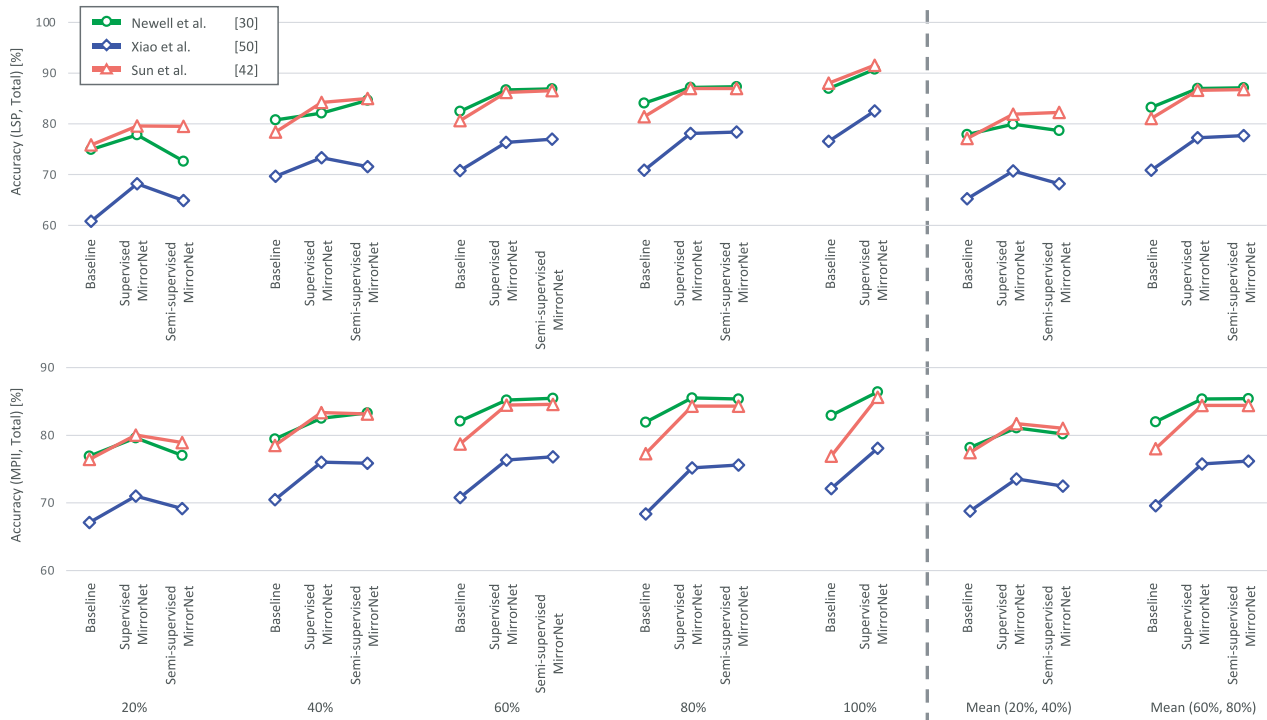


Fig. 9 Pose estimation performance (“Total” columns of Tables 1–6). The top and bottom rows show the respective performance on the LSP dataset [18], [19] and the MPII dataset [1], respectively. From left to right, the first five columns show the performance under the conditions that 20%, 40%, 60%, 80%, and 100% of the training data were regarded as annotated images, respectively, and the last two columns show the average respective performance under the 20% and 40% conditions and those under the 60% and 80% conditions, respectively.

Table 7 Pose estimation performance on the LSP dataset [18], [19] with the pose recognizer α based on HRNet [42] with respect to the number of training epochs.

	Training epochs	Training data		PCK@0.2							
		#annotated	#non-annotated	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Supervised MirrorNet	200	8800	-	95.61	89.69	85.21	83.90	83.98	85.31	83.67	86.98
Semi-supervised MirrorNet	200	8800	2200	95.70	89.91	85.52	83.60	82.96	85.78	83.81	86.95
Supervised MirrorNet	300	8800	-	96.26	89.83	86.29	84.85	84.20	85.79	83.81	87.47
Semi-supervised MirrorNet	300	8800	2200	95.92	89.85	86.40	84.81	84.50	86.52	83.81	87.61

Table 8 Pose estimation performance on the MPII dataset [1] with the pose recognizer α based on HGNet [30] with respect to the number of training epochs.

	Training epochs	Training data		PCK@0.2							
		#annotated	#non-annotated	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Supervised MirrorNet	200	17797	-	95.85	93.45	85.40	79.32	85.24	79.32	75.69	85.54
Semi-supervised MirrorNet	200	17797	4449	95.76	93.36	85.14	79.21	84.79	79.29	75.64	85.38
Supervised MirrorNet	300	17797	-	95.96	93.46	85.91	80.13	85.39	80.34	76.70	86.02
Semi-supervised MirrorNet	300	17797	4449	95.92	93.67	86.08	80.17	85.73	80.42	76.73	86.15

method. This strongly supports hypothesis (A) or namely that the joint training of the generative and recognition models leads to performance improvement. The fidelity and plausibility of estimated poses, which were evaluated by the pose-to-image generator θ and the pose VAE, respectively, were key factors for accurate pose estimation.

We found that the semi-supervised MirrorNet outperformed the supervised MirrorNet when the ratios of annotated images were higher in the training data. As shown in the rightmost column of Fig. 9, the semi-supervised MirrorNet outperformed the

supervised MirrorNet when the annotation ratio was 60% or 80%, except for the conditions that the HRNet [42] and the HGNet [30] were used as the pose recognizer α on the LSP dataset and the MPII dataset, respectively. When the number of epochs was increased from 50 to 100 in each of the steps (2) and (3) on trial, the semi-supervised MirrorNet performed better under those conditions, as shown in Tables 7 and 8. The performance had often been degraded temporarily before the pose recognizer α learned to extract consistent representations from annotated and non-annotated images. As shown in the second right column



Fig. 10 Examples of pose estimation obtained by the baseline method [30], the supervised and semi-supervised versions of MirrorNet. Anatomically implausible poses were corrected by the MirrorNet architecture.

Table 9 Network size and computation speed.

Network	#params	GFLOPs (LSP)	GFLOPs (MPII)
Pose recognizer α			
– Hourglass [30]	25.59M	26.17	19.62
– ResNet-50 [50]	34.00M	11.99	8.99
– HRNet [42]	28.54M	9.49	7.12
Pose-conditioned image VAE			
– Appear. & scene recognizers β & γ	5.28M	1.11	0.83
– Image generator θ	12.84M	3.25	2.44
– Mask estimator ψ	10.28M	1.42	1.06
Pose VAE			
– Primitive recognizer δ	5.29M	1.13	0.85
– Pose generator ϕ	1.33M	1.13	0.85
MirrorNet (training)			
– α (Hourglass [30]), β , γ , δ , θ , ϕ , & ψ	65.89M	35.32	26.48
– α (ResNet-50 [50]), β , γ , δ , θ , ϕ , & ψ	74.30M	21.14	15.85
– α (HRNet [42]), β , γ , δ , θ , ϕ , & ψ	68.84M	18.64	13.98
MirrorNet (runtime)			
– only α (Hourglass [30])	25.59M	26.17	19.62
– only α (ResNet-50 [50])	34.00M	11.99	8.99
– only α (HRNet [42])	28.54M	9.49	7.12

of Fig. 9, in contrast, the semi-supervised MirrorNet underperformed the supervised MirrorNet when the annotation ratio was 20% or 40%. Under these conditions, the pose-to-image generator θ and the pose VAE could not appropriately evaluate the fidelity and plausibility of estimated poses, i.e., gave wrong feedback to the pose recognizer α in steps (2) and (3), leading to the performance degradation. These results conditionally support hypothesis (B) or namely that the semi-supervised training is effective under the condition that a sufficient number of annotated images are available and MirrorNet is sufficiently trained in the semi-supervised fine-tuning steps.

As shown in **Fig. 10**, the pose recognizer α trained by using the MirrorNet architecture yielded anatomically plausible poses. For a better understanding of how each part of MirrorNet works, we show examples of human images generated by the pose-conditioned VAE in Fig. A-1, pose images by the pose VAE in Fig. A-2, and silhouette images by the mask estimator ϕ in Fig. A-3 in the appendix. As shown in **Table 9**, the training of the whole MirrorNet is computationally demanding because the generative and recognition models of pose and images should be trained jointly. Note that only the pose recognizer α is used in

the runtime; the pose-conditioned VAE and the pose VAE serve as regularizers that stabilize the training of the MirrorNet.

5.4 Discussions

Effectiveness of Individual Components. To validate the effectiveness of each component of MirrorNet, we conducted an ablation study. We used the semi-supervised MirrorNet with the HRNet-based pose recognizer α [42] trained on the LSP dataset, where the ratio of annotated images was 20%, 40%, 60%, or 80%. **Tables 10** and **11** show the pose estimation performance obtained with and without the lower-level mirror system and the mask estimator. This clearly shows the effectiveness of the lower-level mirror system and the mask estimator, respectively. Figure A-4 shows examples of human images generated by the pose-conditioned VAE with and without the mask estimator. These results strongly support the design of MirrorNet.

Effectiveness of Curriculum Learning. To validate the effectiveness of the curriculum learning consisting of the pre-training of the individual sub-networks and the fine-tuning of the whole network, we conducted another additional experiment. We used the semi-supervised MirrorNet with the HRNet-based pose recognizer α [42] trained on the LSP dataset, where the ratio of annotated images was 20%, 40%, 60%, or 80%. **Table 12** shows the pose estimation performance obtained with and without the pre-training. When the ratio of annotated images was larger, MirrorNet without the pre-training outperformed that with the pre-training, i.e., training the whole network from scratch is more effective than the stability-aware gradual optimization for finding a better solution. It is thus important to select an appropriate training procedure for extracting the full potential of the semi-supervised MirrorNet.

Effectiveness of Semi-supervised Learning. To investigate the impact of the mini-batch composition in the semi-supervised training of MirrorNet, we changed the number of annotated and non-annotated images in each mini-batch to 32+96, 48+80, 64+64, 80+48, or 96+32. We used MirrorNet with the HRNet-based pose recognizer α [42] trained on the LSP dataset, where the ratio of annotated images was 20%, 40%, 60%, or 80%. As shown in Section 5.3, the semi-supervised MirrorNet underper-

Table 10 Pose estimation performance obtained with and without the lower-level mirror system.

	Training data		PCK@0.2							
	#annotated	#non-annotated	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
MirrorNet w/o pose VAE	2200	8800	94.43	84.02	76.25	75.09	75.64	72.62	69.10	78.31
MirrorNet w/ pose VAE	2200	8800	92.55	84.16	77.90	76.58	73.57	76.14	74.06	79.49
MirrorNet w/o pose VAE	4400	6600	94.85	86.63	80.74	78.57	79.95	78.80	76.49	82.49
MirrorNet w/ pose VAE	4400	6600	94.47	88.48	83.31	82.57	81.63	82.95	80.23	85.02
MirrorNet w/o pose VAE	6600	4400	95.58	87.75	82.95	81.07	81.32	83.08	80.12	84.77
MirrorNet w/ pose VAE	6600	4400	95.69	89.40	85.11	83.37	83.28	84.58	83.23	86.58
MirrorNet w/o pose VAE	8800	2200	95.46	88.50	83.53	81.33	83.09	83.30	80.96	85.35
MirrorNet w/ pose VAE	8800	2200	95.70	89.91	85.52	83.60	82.96	85.78	83.81	86.95
MirrorNet w/o pose VAE	11000	-	96.98	92.51	89.30	87.89	89.28	90.35	88.38	90.81
MirrorNet w/ pose VAE	11000	-	96.65	92.78	89.71	87.44	91.72	91.57	89.90	91.58

Table 11 Pose estimation performance obtained with and without the mask estimator.

	Training data		PCK@0.2							
	#annotated	#non-annotated	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
MirrorNet w/o mask estimator	2200	8800	91.62	83.85	77.15	76.38	72.35	74.77	72.49	78.62
MirrorNet w/ mask estimator	2200	8800	92.55	84.16	77.90	76.58	73.57	76.14	74.06	79.49
MirrorNet w/o mask estimator	4400	6600	94.58	88.58	83.52	81.83	81.10	81.77	79.77	84.64
MirrorNet w/ mask estimator	4400	6600	94.47	88.48	83.31	82.57	81.63	82.95	80.23	85.02
MirrorNet w/o mask estimator	6600	4400	95.43	88.60	83.53	81.53	82.77	83.68	81.62	85.49
MirrorNet w/ mask estimator	6600	4400	95.69	89.40	85.11	83.37	83.28	84.58	83.23	86.58
MirrorNet w/o mask estimator	8800	2200	95.55	88.70	84.86	83.63	82.81	85.00	83.45	86.47
MirrorNet w/ mask estimator	8800	2200	95.70	89.91	85.52	83.60	82.96	85.78	83.81	86.95
MirrorNet w/o mask estimator	11000	-	96.91	92.55	89.06	87.79	90.00	90.22	87.53	90.68
MirrorNet w/ mask estimator	11000	-	96.65	92.78	89.71	87.44	91.72	91.57	89.90	91.58

Table 12 Pose estimation performance obtained with and without the supervised pre-training.

	Training data		PCK@0.2							
	#annotated	#non-annotated	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
MirrorNet w/o supervised pre-training	2200	8800	91.52	82.02	76.37	74.79	73.76	75.15	72.32	78.22
MirrorNet w/ supervised pre-training	2200	8800	92.55	84.16	77.90	76.58	73.57	76.14	74.06	79.49
MirrorNet w/o supervised pre-training	4400	6600	93.54	87.31	82.40	81.38	77.93	81.80	79.11	83.56
MirrorNet w/ supervised pre-training	4400	6600	94.47	88.48	83.31	82.57	81.63	82.95	80.23	85.02
MirrorNet w/o supervised pre-training	6600	4400	95.16	88.84	85.53	84.47	81.82	85.63	83.60	86.61
MirrorNet w/ supervised pre-training	6600	4400	95.69	89.40	85.11	83.37	83.28	84.58	83.23	86.58
MirrorNet w/o supervised pre-training	8800	2200	95.85	91.27	87.68	85.63	85.63	88.09	86.08	88.82
MirrorNet w/ supervised pre-training	8800	2200	95.70	89.91	85.52	83.60	82.96	85.78	83.81	86.95
MirrorNet w/o supervised pre-training	11000	-	97.05	93.59	91.51	90.63	91.99	92.73	91.35	92.79
MirrorNet w/ supervised pre-training	11000	-	96.65	92.78	89.71	87.44	91.72	91.57	89.90	91.58

formed the supervised MirrorNet when the mini-batch size was 96+32 and the ratio of annotated images was 20%. Interestingly, as shown in **Table 13**, the semi-supervised MirrorNet outperformed the supervised MirrorNet under the conditions of 32+96 and 64+64. In the objective function given by Eq. (29), the contributions of annotated and non-annotated images are directly affected by the ratio of annotated images in each mini-batch. Thus, it is necessary to optimize it for drawing out the full potential of semi-supervised learning. This should be included as topics for future work.

Future Directions. One of the most interesting research direc-

tions is to investigate fully unsupervised training of MirrorNet because it is based on the VAE architecture and can be trained from only non-annotated images in theory. In this study, the hierarchical mirror system in single-person 2D pose estimation has successfully been used only under the semi-supervised condition. Another important research direction is to deal with human images in which some joints are occluded or out of view. The noticeable advantage of the fully probabilistic modeling underlying MirrorNet is that unobserved joints could be statistically inferred during the training. In addition, it is worth extending the current MirrorNet for multi-person 3D pose estimation by using a hier-

Table 13 Pose estimation performance with respect to the ratio of annotated images in each mini-batch.

	Training data		Mini-batch composition		PCK@0.2							
	#annotated	#non-annotated	#annotated	#non-annotated	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Supervised MirrorNet	2200	-	128	-	94.04	83.92	78.19	76.61	75.90	74.82	72.30	79.61
Semi-supervised MirrorNet	2200	8800	32	96	94.00	84.86	78.71	77.33	78.07	77.21	74.88	80.95
	2200	8800	48	80	94.08	82.55	78.11	76.86	73.89	74.89	72.97	79.28
	2200	8800	64	64	93.78	84.89	79.02	77.10	77.28	76.81	73.91	80.61
	2200	8800	80	48	92.67	84.45	78.83	76.82	75.38	75.29	72.27	79.59
	2200	8800	96	32	92.55	84.16	77.90	76.58	73.57	76.14	74.06	79.49
Supervised MirrorNet	4400	-	128	-	95.02	88.23	82.45	81.23	81.30	81.16	78.76	84.22
Semi-supervised MirrorNet	4400	6600	32	96	94.67	87.71	80.63	79.07	80.92	80.19	77.40	83.14
	4400	6600	48	80	95.20	87.86	81.76	80.27	81.92	81.26	78.05	83.94
	4400	6600	64	64	95.07	88.84	83.20	81.09	82.33	82.11	79.63	84.82
	4400	6600	80	48	94.94	88.30	81.98	80.85	81.75	83.58	80.72	84.82
	4400	6600	96	32	94.47	88.48	83.31	82.57	81.63	82.95	80.23	85.02
Supervised MirrorNet	6600	-	128	-	95.65	88.82	84.34	82.74	83.42	84.74	82.36	86.22
Semi-supervised MirrorNet	6600	4400	32	96	95.08	86.98	80.41	78.71	80.76	80.81	77.59	83.11
	6600	4400	48	80	95.93	89.18	84.10	81.73	82.45	83.38	80.61	85.50
	6600	4400	64	64	95.37	88.21	82.98	82.00	81.76	84.08	81.93	85.43
	6600	4400	80	48	95.71	88.62	84.18	82.69	83.02	84.68	82.67	86.15
	6600	4400	96	32	95.69	89.40	85.11	83.37	83.28	84.58	83.23	86.58
Supervised MirrorNet	8800	-	128	-	95.61	89.69	85.21	83.90	83.98	85.31	83.67	86.98
Semi-supervised MirrorNet	8800	2200	32	96	94.39	86.12	80.50	78.95	78.67	78.05	76.34	82.05
	8800	2200	48	80	94.60	88.22	83.22	81.42	82.54	81.83	80.14	84.78
	8800	2200	64	64	94.96	88.47	83.47	82.08	82.87	83.64	82.06	85.58
	8800	2200	80	48	95.44	89.62	85.17	83.25	83.55	84.66	82.76	86.56
	8800	2200	96	32	95.70	89.91	85.52	83.60	82.96	85.78	83.81	86.95

archical mirror system involving the 3D pose VAE at the lower level.

6. Conclusion

Inspired by the cognitive knowledge about the mirror neuron system of humans, this paper proposes a deep Bayesian framework called MirrorNet for 2D pose estimation from human images. The key idea is to jointly train the generative models of images and poses as well as the recognition models of appearances, scenes, and primitives in a fully statistical manner. From a technical point of view, the two-level mirror systems (VAEs) are jointly trained with the hierarchical autoencoding manner (image \rightarrow pose \rightarrow primitive \rightarrow pose \rightarrow image), such that the plausibility and fidelity of poses are both considered. Thanks to the nature of the fully generative modeling, MirrorNet is the first pose estimation architecture that could, in theory, be trained from non-annotated images in an unsupervised manner when appropriate inductive biases are introduced. We experimentally proved that the whole MirrorNet could be jointly trained and outperformed a conventional recognition-model-only method in terms of pose estimation performance. We also showed that the additional use of non-annotated images could improve the performance.

The main contribution of this paper is that we shed light on the mirror neuron system (or motor theory) and built a robust computational model of the human vision system by leveraging the expressive power of modern deep Bayesian models. The same framework can be applied to 3D motion estimation from videos by formulating recurrent versions of the pose and image VAEs that represent the anatomical plausibility and fidelity of human

motions, respectively. This paper also ushers the new research field of semi-supervised pose estimation. We believe that MirrorNet inspires a new approach to multimedia understanding.

Acknowledgments We are thankful for AI Bridging Cloud Infrastructure (ABCI) of National Institute of Advanced Industrial Science and Technology (AIST), which we used extensively for our experiments. This work was partly supported by the Program for Leading Graduate Schools, “Graduate Program for Embodiment Informatics” of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, JST ACCEL No.JPMJAC1602, JSPS KAKENHI No.19H04137, and JST-Mirai Program No.JPMJMI19B2.

References

- [1] Andriluka, M., Pishchulin, L., Gehler, P. and Schiele, B.: 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3686–3693 (2014).
- [2] Andriluka, M., Roth, S. and Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1014–1021 (2009).
- [3] Belagiannis, V. and Zisserman, A.: Recurrent human pose estimation. *International Conference on Automatic Face & Gesture Recognition (FG)*, pp.468–475 (2017).
- [4] Bulat, A. and Tzimiropoulos, G.: Human pose estimation via convolutional part heatmap regression. *European Conference on Computer Vision (ECCV)*, pp.717–732 (2016).
- [5] Carreira, J., Agrawal, P., Fragkiadaki, K. and Malik, J.: Human pose estimation with iterative error feedback. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.4733–4742 (2016).
- [6] Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G. and Sun, J.: Cascaded pyramid network for multi-person pose estimation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.7103–7112 (2018).
- [7] Chen, Y., Shen, C., Wei, X.-S., Liu, L. and Yang, J.: Adversarial

- PoseNet: A structure-aware convolutional network for human pose estimation, *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1212–1221 (2017).
- [8] Chou, C.-J., Chien, J.-T. and Chen, H.-T.: Self adversarial training for human pose estimation, *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp.17–30 (2018).
- [9] Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L. and Wang, X.: Multi-context attention for human pose estimation, *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1831–1840 (2017).
- [10] Dai, Z., Damianou, A., Gonzalez, J. and Lawrence, N.: Variational Auto-encoded Deep Gaussian Processes, *International Conference on Learning Representations (ICLR)*, pp.1–11 (2016).
- [11] Dantone, M., Gall, J., Leistner, C. and Van Gool, L.: Human pose estimation using body parts dependent joint regressors, *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3041–3048 (2013).
- [12] de Bem, R., Ghosh, A., Ajanthan, T., Miksik, O., Boukhayma, A., Siddharth, N. and Torr, P.: DGPose: Deep Generative Models for Human Body Analysis, *International Journal of Computer Vision (IJCV)*, Vol.128, pp.1537–1563 (2020).
- [13] Fieraru, M., Khoreva, A., Pishchulin, L. and Schiele, B.: Learning to refine human pose estimation, *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp.205–214 (2018).
- [14] Gkioxari, G., Arbelaez, P., Bourdev, L. and Malik, J.: Articulated pose estimation using discriminative armllet classifiers, *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3342–3349 (2013).
- [15] Hawking, S.: *The Universe in a Nutshell. The Inspiring Sequel to A Brief History of Time*, London: Transworld Publishers (2001).
- [16] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770–778 (2016).
- [17] Iacono, M. and Mazziotta, J.C.: Mirror Neuron System: Basic Findings and Clinical Applications, *American Neurological Association*, Vol.62, No.3, pp.213–218 (2007).
- [18] Johnson, S. and Everingham, M.: Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation, *British Machine Vision Conference (BMVC)*, pp.1–11 (2010).
- [19] Johnson, S. and Everingham, M.: Learning effective human pose estimation from inaccurate annotation, *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1465–1472 (2011).
- [20] Ke, L., Chang, M.-C., Qi, H. and Lyu, S.: Multi-scale structure-aware network for human pose estimation, *European Conference on Computer Vision (ECCV)*, pp.713–728 (2018).
- [21] Kim, T., Cha, M., Kim, H., Lee, J.K. and Kim, J.: Learning to discover cross-domain relations with generative adversarial networks, *International Conference on Machine Learning (ICML)*, pp.1857–1865 (2017).
- [22] Kingma, D.P. and Ba, J.: Adam: A Method for Stochastic Optimization, *International Conference for Learning Representations (ICLR)*, pp.1–13 (2015).
- [23] Kingma, D.P., Rezende, D.J., Mohamed, S. and Welling, M.: Semi-supervised Learning with Deep Generative Models, *Advances in Neural Information Processing Systems (NIPS)*, pp.3581–3589 (2014).
- [24] Kingma, D.P. and Welling, M.: Auto-Encoding Variational Bayes, *International Conference on Learning Representations (ICLR)*, pp.1–14 (2014).
- [25] Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J. and Gehler, P.V.: Unite the People: Closing the Loop Between 3D and 2D Human Representations, *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.6050–6059 (2017).
- [26] Lifshitz, I., Fetaya, E. and Ullman, S.: Human pose estimation using deep consensus voting, *European Conference on Computer Vision (ECCV)*, pp.246–260 (2016).
- [27] Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B. and Fritz, M.: Disentangled person image generation, *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.99–108 (2018).
- [28] Mnih, A. and Gregor, K.: Neural Variational Inference and Learning in Belief Networks, *International Conference on Machine Learning (ICML)*, pp.1791–1799 (2014).
- [29] Moon, G., Chang, J.Y. and Lee, K.M.: Posefix: Model-agnostic general human pose refinement network, *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.7773–7781 (2019).
- [30] Newell, A., Yang, K. and Deng, J.: Stacked hourglass networks for human pose estimation, *European Conference on Computer Vision (ECCV)*, pp.483–499 (2016).
- [31] Nie, X., Feng, J., Zhang, J. and Yan, S.: Single-stage multi-person pose machines, *International Conference on Computer Vision (ICCV)*, pp.6951–6960 (2019).
- [32] Nie, X., Feng, J., Zuo, Y. and Yan, S.: Human pose estimation with parsing induced learner, *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.2100–2108 (2018).
- [33] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: PyTorch: An imperative style, high-performance deep learning library, *Advances in Neural Information Processing Systems (NeurIPS)*, pp.8024–8035 (2019).
- [34] Peng, X., Tang, Z., Yang, F., Feris, R.S. and Metaxas, D.: Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation, *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.2226–2234 (2018).
- [35] Pishchulin, L., Andriluka, M., Gehler, P. and Schiele, B.: Poselet conditioned pictorial structures, *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.588–595 (2013).
- [36] Qiao, T., Zhang, J., Xu, D. and Tao, D.: MirrorGAN: Learning Text-to-image Generation by Redescription, *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1505–1514 (2019).
- [37] Ramanan, D.: Learning to parse images of articulated bodies, *Advances in Neural Information Processing Systems (NIPS)*, pp.1129–1136 (2007).
- [38] Rezende, D.J., Mohamed, S. and Wierstra, D.: Stochastic Backpropagation and Approximate Inference in Deep Generative Models, *International Conference on Machine Learning (ICML)*, pp.1278–1286 (2014).
- [39] Ronneberger, O., Fischer, P. and Brox, T.: U-Net: Convolutional networks for biomedical image segmentation, *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, pp.234–241 (2015).
- [40] Sapp, B. and Taskar, B.: Modec: Multimodal decomposable models for human pose estimation, *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3674–3681 (2013).
- [41] Sapp, B., Jordan, C. and Taskar, B.: Adaptive pose priors for pictorial structures, *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.422–429 (2010).
- [42] Sun, K., Xiao, B., Liu, D. and Wang, J.: Deep High-Resolution Representation Learning for Human Pose Estimation, *Computer Vision and Pattern Recognition (CVPR)*, pp.5693–5703 (2019).
- [43] Tang, W., Yu, P. and Wu, Y.: Deeply learned compositional models for human pose estimation, *European Conference on Computer Vision (ECCV)*, pp.190–206 (2018).
- [44] Tompson, J.J., Jain, A., LeCun, Y. and Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation, *Advances in Neural Information Processing Systems (NIPS)*, pp.1799–1807 (2014).
- [45] Toshev, A. and Szegedy, C.: DeepPose: Human pose estimation via deep neural networks, *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1653–1660 (2014).
- [46] Ukita, N. and Uematsu, Y.: Semi- and weakly-supervised human pose estimation, *Computer Vision and Image Understanding*, Vol.170, pp.67–78 (2018).
- [47] Walker, J., Marino, K., Gupta, A. and Hebert, M.: The pose knows: Video forecasting by generating pose futures, *International Conference on Computer Vision (ICCV)*, pp.3332–3341 (2017).
- [48] Wei, S.-E., Ramakrishna, V., Kanade, T. and Sheikh, Y.: Convolutional pose machines, *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.4724–4732 (2016).
- [49] Xiao, B.: An official implementation of “Deep High-Resolution Representation Learning for Human Pose Estimation”, available from <https://github.com/leoxiaobin/deep-high-resolution-net.pytorch> (accessed 2020-03-17).
- [50] Xiao, B., Wu, H. and Wei, Y.: Simple baselines for human pose estimation and tracking, *European Conference on Computer Vision (ECCV)*, pp.466–481 (2018).
- [51] Yang, W., Li, S., Ouyang, W., Li, H. and Wang, X.: Learning feature pyramids for human pose estimation, *International Conference on Computer Vision (ICCV)*, pp.1281–1290 (2017).
- [52] Yang, Y. and Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts, *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1385–1392 (2011).
- [53] Yang, Y. and Ramanan, D.: Articulated human detection with flexible mixtures of parts, *IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI)*, Vol.35, No.12, pp.2878–2890 (2013).
- [54] Yeh, R.A., Hu, Y.-T. and Schwing, A.G.: Chirality Nets: Exploiting Structure in Human Pose Regression, *Conference on Advances in Neural Information Processing Systems Workshop (NeurIPSWS)*, pp.1–11 (2019).
- [55] Yi, Z., Zhang, H., Tan, P. and Gong, M.: DualGAN: Unsupervised dual learning for image-to-image translation, *International Conference on Computer Vision (ICCV)*, pp.2849–2857 (2017).
- [56] Yildirim, I., Belledonne, M., Freiwald, W. and Tenenbaum, J.: Effi-

cient inverse graphics in biological face processing, *Science Advances*, Vol.6, No.10, pp.1–18 (2020).

- [57] Zhu, J.-Y., Park, T., Isola, P. and Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks, *International Conference on Computer Vision (ICCV)*, pp.2223–2232 (2017).

Appendix

A.1 Lower Bound \mathcal{L}

The variational lower bound \mathcal{L}^X of $\log p(\mathbf{X})$ in the unsupervised condition ($\mathcal{L}^X = \sum_n \mathcal{L}_n^X$) is given by Eq. (A.1). The variational lower bound $\mathcal{L}^{X,S}$ of $\log p(\mathbf{X}, \mathbf{S})$ in the supervised condition ($\mathcal{L}^{X,S} = \sum_n \mathcal{L}_n^{X,S}$) is given by Eq. (A.2).

$$\begin{aligned}
\mathcal{L}_n^X &= \mathbb{E}_{q(\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n, \mathbf{z}_n | \mathbf{x}_n)} [\log \{ p_\theta(\mathbf{x}_n | \mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n) p_\phi(\mathbf{s}_n | \mathbf{z}_n) p(\mathbf{a}_n) p(\mathbf{g}_n) p(\mathbf{z}_n) \} \\
&\quad - \log \{ q_\alpha(\mathbf{s}_n | \mathbf{x}_n) q_\beta(\mathbf{a}_n | \mathbf{s}_n, \mathbf{x}_n) q_\gamma(\mathbf{g}_n | \mathbf{s}_n, \mathbf{x}_n) q_\delta(\mathbf{z}_n | \mathbf{s}_n) \}] \\
&= \mathbb{E}_{q_\alpha(\mathbf{s}_n | \mathbf{x}_n) q_\beta(\mathbf{a}_n | \mathbf{s}_n, \mathbf{x}_n) q_\gamma(\mathbf{g}_n | \mathbf{s}_n, \mathbf{x}_n)} [\log p_\theta(\mathbf{x}_n | \mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n)] \\
&\quad + \mathbb{E}_{q_\alpha(\mathbf{s}_n | \mathbf{x}_n) q_\delta(\mathbf{z}_n | \mathbf{s}_n)} [\log p_\phi(\mathbf{s}_n | \mathbf{z}_n)] - \mathbb{E}_{q_\alpha(\mathbf{s}_n | \mathbf{x}_n)} [\log q_\alpha(\mathbf{s}_n | \mathbf{x}_n)] \\
&\quad - \mathbb{E}_{q_\alpha(\mathbf{s}_n | \mathbf{x}_n)} [\text{KL}(q_\beta(\mathbf{a}_n | \mathbf{s}_n, \mathbf{x}_n) \| p(\mathbf{a}_n))] \\
&\quad - \mathbb{E}_{q_\alpha(\mathbf{s}_n | \mathbf{x}_n)} [\text{KL}(q_\gamma(\mathbf{g}_n | \mathbf{s}_n, \mathbf{x}_n) \| p(\mathbf{g}_n))] \\
&\quad - \mathbb{E}_{q_\alpha(\mathbf{s}_n | \mathbf{x}_n)} [\text{KL}(q_\delta(\mathbf{z}_n | \mathbf{s}_n) \| p(\mathbf{z}_n))] \\
&= -\frac{1}{2} \sum_{d_x=1}^{D^x} \mathbb{E}_{q_\alpha(\mathbf{s}_n | \mathbf{x}_n) q_\beta(\mathbf{a}_n | \mathbf{s}_n, \mathbf{x}_n) q_\gamma(\mathbf{g}_n | \mathbf{s}_n, \mathbf{x}_n)} \left[\log(2\pi\sigma_{\theta, d_x}^2(\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n)) \right. \\
&\quad \left. + \frac{(\mathbf{x}_n - \mu_{\theta, d_x}(\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n))^2}{\sigma_{\theta, d_x}^2(\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n)} \right] \\
&\quad - \frac{1}{2} \sum_{d_s=1}^{D^s} \mathbb{E}_{q_\alpha(\mathbf{s}_n | \mathbf{x}_n) q_\delta(\mathbf{z}_n | \mathbf{s}_n)} \left[\log(2\pi\sigma_{\phi, d_s}^2(\mathbf{z}_n)) + \frac{(\mathbf{s}_n - \mu_{\phi, d_s}(\mathbf{z}_n))^2}{\sigma_{\phi, d_s}^2(\mathbf{z}_n)} \right] \\
&\quad + \frac{1}{2} \sum_{d_s=1}^{D^s} (1 + \log(2\pi\sigma_{\alpha, d_s}^2(\mathbf{x}_n))) \\
&\quad + \frac{1}{2} \sum_{d_a=1}^{D^a} \mathbb{E}_{q_\alpha(\mathbf{s}_n | \mathbf{x}_n)} [1 + \log(\sigma_{\beta, d_a}^2(\mathbf{s}_n, \mathbf{x}_n)) \\
&\quad \quad - \mu_{\beta, d_a}^2(\mathbf{s}_n, \mathbf{x}_n) - \sigma_{\beta, d_a}^2(\mathbf{s}_n, \mathbf{x}_n)] \\
&\quad + \frac{1}{2} \sum_{d_g=1}^{D^g} \mathbb{E}_{q_\alpha(\mathbf{s}_n | \mathbf{x}_n)} [1 + \log(\sigma_{\gamma, d_g}^2(\mathbf{s}_n, \mathbf{x}_n)) \\
&\quad \quad - \mu_{\gamma, d_g}^2(\mathbf{s}_n, \mathbf{x}_n) - \sigma_{\gamma, d_g}^2(\mathbf{s}_n, \mathbf{x}_n)] \\
&\quad + \frac{1}{2} \sum_{d_z=1}^{D^z} \mathbb{E}_{q_\alpha(\mathbf{s}_n | \mathbf{x}_n)} [1 + \log(\sigma_{\delta, d_z}^2(\mathbf{s}_n)) \\
&\quad \quad - \mu_{\delta, d_z}^2(\mathbf{s}_n) - \sigma_{\delta, d_z}^2(\mathbf{s}_n)]. \tag{A.1}
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}_n^{X,S} &= \mathbb{E}_{q(\mathbf{a}_n, \mathbf{g}_n, \mathbf{z}_n | \mathbf{s}_n, \mathbf{x}_n)} [\log \{ p_\theta(\mathbf{x}_n | \mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n) p_\phi(\mathbf{s}_n | \mathbf{z}_n) \\
&\quad \times p(\mathbf{a}_n) p(\mathbf{g}_n) p(\mathbf{z}_n) \} \\
&\quad - \log \{ q_\beta(\mathbf{a}_n | \mathbf{s}_n, \mathbf{x}_n) q_\gamma(\mathbf{g}_n | \mathbf{s}_n, \mathbf{x}_n) q_\delta(\mathbf{z}_n | \mathbf{s}_n) \}] \\
&= \mathbb{E}_{q_\beta(\mathbf{a}_n | \mathbf{s}_n, \mathbf{x}_n) q_\gamma(\mathbf{g}_n | \mathbf{s}_n, \mathbf{x}_n)} [\log p_\theta(\mathbf{x}_n | \mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n)] \\
&\quad + \mathbb{E}_{q_\alpha(\mathbf{s}_n | \mathbf{x}_n)} [\log p_\phi(\mathbf{s}_n | \mathbf{z}_n)] \\
&\quad - \text{KL}(q_\beta(\mathbf{a}_n | \mathbf{s}_n, \mathbf{x}_n) \| p(\mathbf{a}_n)) - \text{KL}(q_\gamma(\mathbf{g}_n | \mathbf{s}_n, \mathbf{x}_n) \| p(\mathbf{g}_n)) \\
&\quad - \text{KL}(q_\delta(\mathbf{z}_n | \mathbf{s}_n) \| p(\mathbf{z}_n)) \\
&= -\frac{1}{2} \sum_{d_x=1}^{D^x} \mathbb{E}_{q_\beta(\mathbf{a}_n | \mathbf{s}_n, \mathbf{x}_n) q_\gamma(\mathbf{g}_n | \mathbf{s}_n, \mathbf{x}_n)} \left[\log(2\pi\sigma_{\theta, d_x}^2(\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n)) \right.
\end{aligned}$$

$$\begin{aligned}
&\quad \left. + \frac{(\mathbf{x}_n - \mu_{\theta, d_x}(\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n))^2}{\sigma_{\theta, d_x}^2(\mathbf{s}_n, \mathbf{a}_n, \mathbf{g}_n)} \right] \\
&\quad - \frac{1}{2} \sum_{d_s=1}^{D^s} \mathbb{E}_{q_\delta(\mathbf{z}_n | \mathbf{s}_n)} \left[\log(2\pi\sigma_{\phi, d_s}^2(\mathbf{z}_n)) + \frac{(\mathbf{s}_n - \mu_{\phi, d_s}(\mathbf{z}_n))^2}{\sigma_{\phi, d_s}^2(\mathbf{z}_n)} \right] \\
&\quad + \frac{1}{2} \sum_{d_a=1}^{D^a} (1 + \log(\sigma_{\beta, d_a}^2(\mathbf{s}_n, \mathbf{x}_n)) - \mu_{\beta, d_a}^2(\mathbf{s}_n, \mathbf{x}_n) - \sigma_{\beta, d_a}^2(\mathbf{s}_n, \mathbf{x}_n)) \\
&\quad + \frac{1}{2} \sum_{d_g=1}^{D^g} (1 + \log(\sigma_{\gamma, d_g}^2(\mathbf{s}_n, \mathbf{x}_n)) - \mu_{\gamma, d_g}^2(\mathbf{s}_n, \mathbf{x}_n) - \sigma_{\gamma, d_g}^2(\mathbf{s}_n, \mathbf{x}_n)) \\
&\quad + \frac{1}{2} \sum_{d_z=1}^{D^z} (1 + \log(\sigma_{\delta, d_z}^2(\mathbf{s}_n)) - \mu_{\delta, d_z}^2(\mathbf{s}_n) - \sigma_{\delta, d_z}^2(\mathbf{s}_n)). \tag{A.2}
\end{aligned}$$

A.2 Image Reconstruction and Prediction

Figures A-1, A-2, A-3 show the reconstruction of human images based on the pose-conditioned image VAE, the reconstruction of 16-joint heatmaps based on the pose VAE, and the prediction of silhouette images from 16 joint heatmaps based on the mask estimator ϕ , respectively. Figure A-4 shows the reconstruction of human images based on the pose-conditioned image VAE obtained with and without the mask estimator ϕ .

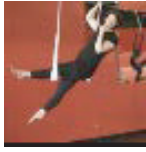
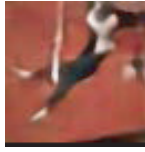
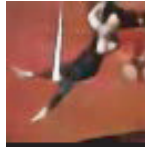
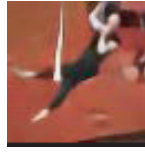
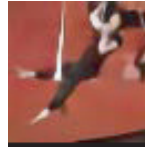
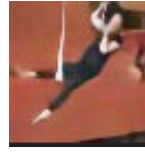
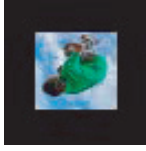
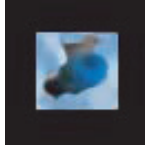
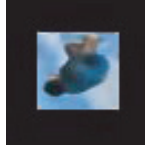
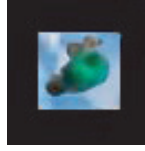
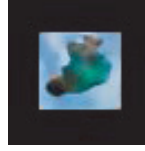
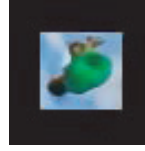

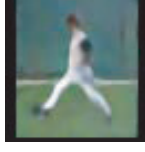
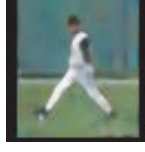
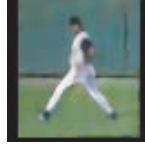
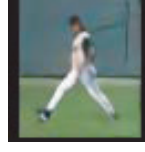
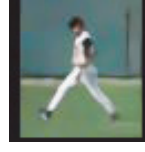
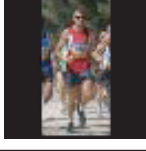
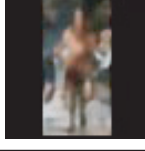
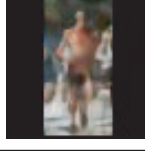
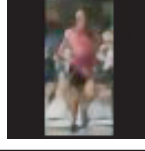
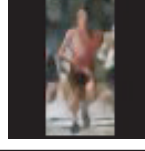
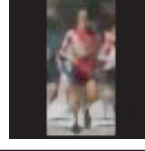
Input	Reconstruction				
	2200	4400	6600	8800	11000
					
					
					
					

Fig. A-1 Reconstruction of human images based on the pose-conditioned image VAE. The LSP dataset [18], [19] was used for training. A larger amount of annotation images resulted in a better quality of generated images.










Original	Input	Reconstruct
		
		
		

Fig. A-2 Reconstruction of 16-joint heatmaps based on the pose VAE. For the purpose of visualization, 16 heatmaps are superimposed to a single heatmap.










Input	Output	Ground truth
		
		
		

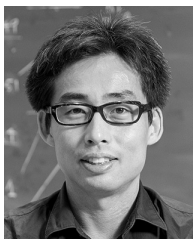
Fig. A-3 Prediction of silhouette images (foreground mask images) from 16 joint heatmaps based on the mask estimator ϕ . The ground truth images are taken from Ref. [25].



Fig. A-4 Reconstruction of human images based on the pose-conditioned image VAE. The LSP dataset [18], [19] was used for training. The ratio of annotated images was 20%.



Takayuki Nakatsuka received a B.E. degree in the Department of Applied Physics and an M.E. degree in the Department of Pure and Applied Physics from Waseda University, Japan. He is currently a student in the Graduate Schools of Advanced Science and Engineering and Graduate Program for Embodiment Informatics at Waseda University. His primary research interests are in physically-based animation, human motion analysis, human computer interaction, and music information retrieval.



Kazuyoshi Yoshii received his M.S. and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2005 and 2008, respectively. He is an Associate Professor with the Graduate School of Informatics, Kyoto University, and concurrently the Leader of the Sound Scene Understanding Team, Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo, Japan. His research interests include music informatics, audio signal processing, and statistical machine learning.



Yuki Koyama is a Researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Japan. He received his Ph.D. from the University of Tokyo in 2017. His main research field is the intersection of computer graphics and human-computer interaction. In particular, he is interested in developing computational techniques for enabling new interactions, producing creative artifacts, and enhancing design processes.



Satoru Fukayama received his Ph.D. degree in information science and technology in 2013 from the University of Tokyo. He is currently a senior researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Japan. His interests are in music information retrieval, especially music generation with probabilistic models. He has received awards, including IPSJ Yamashita SIG Research Award, several Best Presentation Awards, and Specially Selected Paper Award from the Information Processing Society of Japan.



Masataka Goto received his Doctor of Engineering degree from Waseda University in 1998. He is currently a Prime Senior Researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Japan. Over the past 28 years he has published more than 270 papers in refereed journals and international conferences and has received 51 awards, including several best paper awards, best presentation awards, the Tenth Japan Academy Medal, and the Tenth JSPS PRIZE. In 2016, as the Research Director he began OngaACCEL Project, a 5-year JST-funded research project (ACCEL) on music technologies.



Shigeo Morishima was born on August 1959. He received his B.S., M.S. and Ph.D. degrees, all in Electrical Engineering from the University of Tokyo, in 1982, 1984, and 1987, respectively. He was a visiting professor of University of Toronto from 1994 to 1995 and an invited researcher of Advanced Telecommunication Research institute from 1999 to 2011. Currently, he is a professor of School of Advanced Science and Engineering, Waseda University. He was a General Chair of ACM VRST2018 and VR/AR adviser of SIGGRAPH ASIA 2018. He received many awards and takes an administration board member of several societies.