Adversarial PoseNet: A Structure-aware Convolutional Network for Human Pose Estimation*

Yu Chen[1] Chunhua Shen[2] Xiu-Shen Wei[3] Lingqiao Liu[2] Jian Yang[1]
[1]Nanjing University of Science and Technology [2]University of Adelaide [3]Nanjing University

Abstract

For human pose estimation in monocular images, joint occlusions and overlapping upon human bodies often result in deviated pose predictions. Under these circumstances, biologically implausible pose predictions may be produced. In contrast, human vision is able to predict poses by exploiting geometric constraints of joint inter-connectivity. To address the problem by incorporating priors about the structure of human bodies, we propose a novel structure-aware convolutional network to implicitly take such priors into account during training of the deep network. Explicit learning of such constraints is typically challenging. Instead, we design discriminators to distinguish the real poses from the fake ones (such as biologically implausible ones). If the pose generator (G) generates results that the discriminator fails to distinguish from real ones, the network successfully learns the priors.

To better capture the structure dependency of human body joints, the generator G is designed in a stacked multi-task manner to predict poses as well as occlusion heatmaps. Then, the pose and occlusion heatmaps are sent to the discriminators to predict the likelihood of the pose being real. Training of the network follows the strategy of conditional Generative Adversarial Networks (GANs). The effectiveness of the proposed network is evaluated on two widely used human pose estimation benchmark datasets. Our approach significantly outperforms the state-of-the-art methods and almost always generates plausible human pose predictions.

1. Introduction

Human pose estimation is a key step in understanding the actions of people in images and videos. Understanding of a person's limb articulation location is very helpful for high-level vision tasks like human tracking, action recognition, and also serves as a fundamental tool in fields such as human-computer interaction applications. It is a challenging task due to high flexibility of body limbs, self and outer occlusion, various camera angles, etc.

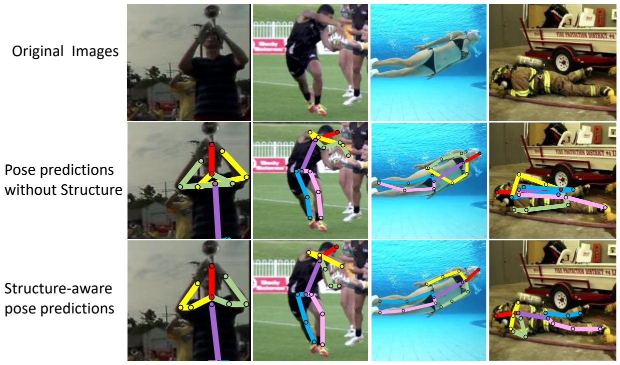
Original Images
Pose predictions without Structure
Structure-aware pose predictions

Figure 1: Motivation. We show the importance of strongly enforcing priors about the human body structure during training of DCNNs for pose estimation. Learning without using such priors generates inaccurate results.

Recently, significant improvements have been achieved on this topic by using Deep Convolutional Neural Networks (DCNNs) [31, 30, 32, 6, 34, 20, 4]. These approaches mainly follow the strategy of regressing heatmaps of body parts using DCNNs. These regression models have shown great ability of learning better feature representations. However, for body parts with heavy occlusions (especially from body parts of surrounding people) and background which seems similar to body parts, DCNNs may have difficulty in regressing accurate heatmaps.

Human vision is capable of learning the variety and limitless of human body shape structures from observations. Even under extreme occlusions, we can infer the potential poses and negative the implausible ones. It is, however, very challenging to incorporate such priors about human body structures into DCNNs, because, as pointed out in [31], the low-level mechanics of DCNNs is typically difficult to interpret, and DCNNs are most capable of learning features.

As a consequence, an unreasonable human pose may be produced by a standard DCNN. As shown in Fig. 1, in challenging test cases with heavy occlusions, standard DCNNs tend to perform poorly. To solve this problem, priors about the structure of the body joints must be considered. The

*This work was done when Y. Chen and X.-S. Wei were visiting University of Adelaide. C. Shen and J. Yang are the corresponding authors.

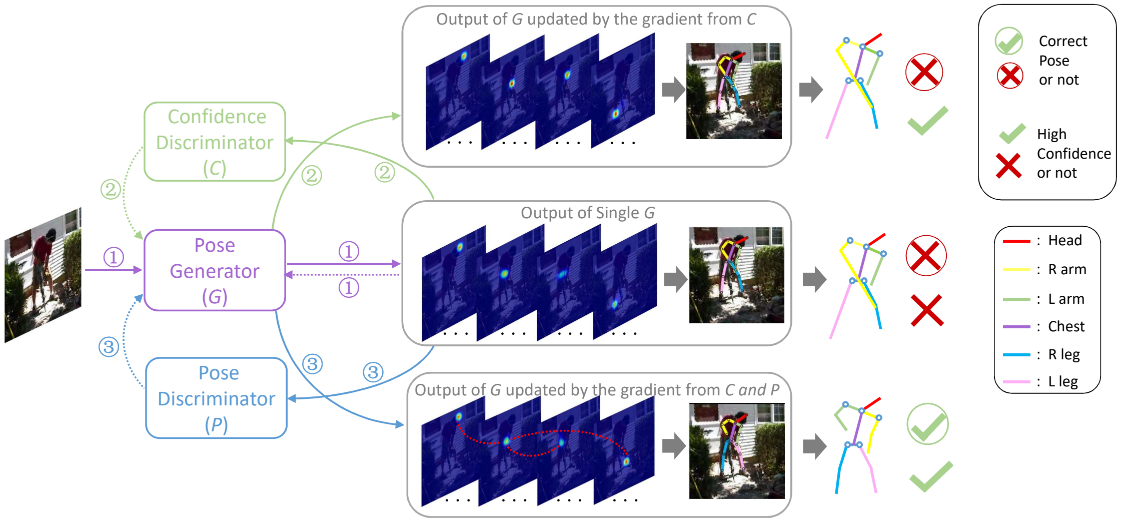

Figure 2: Overview of the proposed Structure-aware Convolutional Network for human pose estimation. The sub-network in purple is the stacked multi-task network (G) for pose generation. The networks in blue (P) and green (C) are used to discriminate whether the generated pose is "real" (reasonable as a body shape) and whether the generator has strong confidence in locating the joints, respectively. Dashed lines into G indicate backward gradients to update G. ① shows the forward and backward of the G net. ② shows the process of G updated by the adversarial gradient from the C net. Then, G is updated by the adversarial gradients from P as shown in lines with ③.

key point of this problem is to learn the real body joints distribution from a large amount of training data. However, explicit learning of such a distribution can be very difficult.

To address this problem, we attempt to learn the distribution of the human body structures *implicitly*. We suppose that we have a "discriminator" which can tell whether the predicted pose is geometrically reasonable. If the DCNN regressor is able to "deceive" the "discriminator" that its predictions are all reasonable, the network would have successfully learned the priors of the human body structure. Inspired by the recent success in Generative Adversarial Networks (GAN) [24, 39, 27, 12, 9], we propose to design the "discriminator" as the discriminator network while the regression network functions as a generative network. Training the generator in the adversarial manner against the discriminator exactly meets our intention.

To accomplish the above goals, the discriminator should be fed with sufficient information to perform classification, while the generator should have the ability in modeling the complicated features in pose estimation. Thus, a multi-task learning network G is designed, which simultaneously regresses the pose heatmaps and the occlusion heatmaps. Based on the pose and occlusion heatmaps, the pose discriminator P is used to tell whether the body configuration is plausible.

In addition, our preliminary results show that correct locations often correspond to highly confident heatmaps. Therefore, we design another discriminator C to make a decision on the confidence of the predicted pose heatmaps. The generator is asked to "fool" both the pose and confidence discriminators by training G and $\{P, C\}$ in the generative adversarial manner. Thus, the human body structure is implied in the P net by guiding G to be close to ground-truth heatmaps and to satisfy joint-connectivity constraints of the human body. The learned G net is expected to be more robust to occlusions and cluttered backgrounds where the precise description for different body parts are required. The main contributions of this work are three folds.

- We design a novel network framework for human pose estimation which takes the geometric constraints of human joints connectivity into consideration. By incorporating the priors of the human body, prediction mistakes caused by occlusions and cluttered backgrounds are considerably reduced. Even when the network fails, the outputs of the network appear more like "human" predictions instead of "machine" predictions.
- To our knowledge, this is the first paper to incorporate adversarial learning into pose estimation; and probably the first to exploit GAN for structured output prediction problems. We also design a stacked multi-task network for predicting both the pose heatmaps and the occlusion heatmaps to achieve better results.
- We evaluate our method on two public human pose estimation datasets. Our approach significantly outperforms state-of-the-art methods, and is able to consistently produce more plausible pose predictions compared to previous methods.

1.1. Related Work

Our work is closely related to work using heatmap based DCNN methods for human pose estimation and Generative Adversarial Networks.

Human Pose Estimation. Traditional human pose estimation methods often follow the framework of tree structured graphical model [10, 3, 29, 36, 22, 28]. With the introduction of "DeepPose" by Toshev *et al.* [32], deep network based methods become more popular in this area. This work is more related to the methods generating pose heatmaps from images [35, 20, 30, 34, 6, 23, 15, 31]. For example, Tompson *et al.* [31] generated heatmaps by running an image through multiple resolution banks in parallel to simultaneously capture features at a variety of scales. Tompson *et al.* [30] used multiple branches of convolutional networks to fuse the features from an image pyramid, and used Markov Random Field (MRF) for post-processing. In the following, Convolutional Pose Machine [34] incorporated the inference of the spatial correlations among body parts within convolutional networks. Hourglass Network [20] proposed a state-of-the-art architecture for bottom-up and top-down inference with residual blocks. The structure of our *G* net is also a fully convolutional network with "conv-deconv" architecture. However, our network is designed in a multi-task manner with features of both tasks delivered into the next stacked network.

Generative Adversarial Network. Generative Adversarial Networks have been widely studied in previous work for discrete labels [19], text [26] and also images. The image-conditional models have tackled inpainting [21], image prediction from a normal map [33], future frame prediction [18], future state prediction [40], product photo generation [38], and style transfer [17]. Human pose estimation can been considered as a translation from a RGB image to a multichannel heatmap. The designed bottom-up and top-down *G* net can well accomplish this translation. Different from previous work, the goal of the discrimination network is not only to distinguish the fake from real, but also to incorporate geometric constraint to the model. This is the reason for the different training strategy for fake samples compared with traditional GANs which will be explained in detail in the following sections.

2. The Proposed Adversarial PoseNet

As mentioned in Fig. 2, our Adversarial PoseNet model consists of three parts, *i.e.*, the pose generator network *G*, the pose discriminator network *P* and the confidence discriminator *C*. The generative network is a bottom-up and top-down network, where the inputs are the RGB images and the outputs are 32 heatmaps for each input image in our case. Half of the returned heatmaps are pose estimations for 16 pose key points, and the other half are for the corresponding occlusion predictions. The values in each heatmap are confidence scores in the range of $[0, 1]$ where a Gaussian blur is done around the ground truth position.

Without discriminators, *G* will be updated simply by forward and backward propagations of itself (cf., the lines with ① in Fig. 2). That might generate low confidence and even incorrect location pose estimations. It is necessary to leverage the power of discriminators to correct these poor estimations. Therefore, two discriminator networks *C* and *P* are introduced into the framework.

After updating *G* by training with *C* in the adversarial manner (cf. the lines with ②), more confident results are produced. Furthermore, after training *G* with both *P* and *C* (cf. the lines with ③), the human body priors are implicitly exploited, and the prediction confidences are accordingly improved.

2.1. Multi-Task Generative Network

In this section, we introduce the generative network *G* in our framework. Fig. 3 presents the architecture of *G*. Knowledge of whether a body part being occluded clearly offers important information for inferring the geometric information of a human pose. Here, in order to effectively incorporate both pose estimation and occlusion predictions, we propose to tackle the problem with a multi-task generative network.

The goal of the multi-task generative network is to learn a function $\mathcal{G}$ which attempts to project an image $\boldsymbol{x}$ to both the corresponding pose heatmaps $\boldsymbol{y}$ and occlusion heatmaps $\boldsymbol{z}$, *i.e.*, $\mathcal{G}(\boldsymbol{x}) = \{\hat{\boldsymbol{y}}, \hat{\boldsymbol{z}}\}$ where $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{z}}$ are the predicted heatmaps. In addition, as reported in [34], large contextual regions are important for locating body parts. So the contextual region of a neuron, which is its receptive field, should be large. To achieve this goal, an "encoder-decoder" architecture is used.

Besides, for the problem of human pose estimation, local evidence is essential for identifying features for human joints. Meanwhile, the final pose estimation requires a coherent understanding of the full body image. To capture this information at each scale, skip connections between mirrored layers in the encoder and decoder are added. Inspired by [20], our network is also stacked to provide the network with a mechanism for re-evaluation of initial estimates and features across the entire image. In each module of the *G* net, a residual block [13] is used for the convolution operator. Given the original image $\boldsymbol{x}$, a basic block of the stacked multi-task generator network can be expressed as follows:

$$\begin{cases} \{\boldsymbol{Y}_n, \boldsymbol{Z}_n, \boldsymbol{X}\} = \mathcal{G}_n(\boldsymbol{Y}_{n-1}, \boldsymbol{Z}_{n-1}, \boldsymbol{X}) & \text{if } n \geqslant 2 \\ \{\boldsymbol{Y}_n, \boldsymbol{Z}_n, \boldsymbol{X}\} = \mathcal{G}_n(\boldsymbol{X}) & \text{if } n = 1 \end{cases},$$

where $\boldsymbol{Y}_n$ and $\boldsymbol{Z}_n$ are the output activation tensors of the n-th stacked generative network for pose estimations and occlusion predictions, respectively. $\boldsymbol{X}$ is the image feature tensor, obtained after pre-processing on the original image through two residual blocks. Suppose that there are N times stacking of the basic block, then the multi-task generative network can be formulated as:

$$\{\boldsymbol{Y}_N, \boldsymbol{Z}_N, \boldsymbol{X}\} = \mathcal{G}_N(\mathcal{G}_{N-1}(\cdots(\mathcal{G}_1(\boldsymbol{X}), \boldsymbol{Y}_1, \boldsymbol{Z}_1)))\,.$$

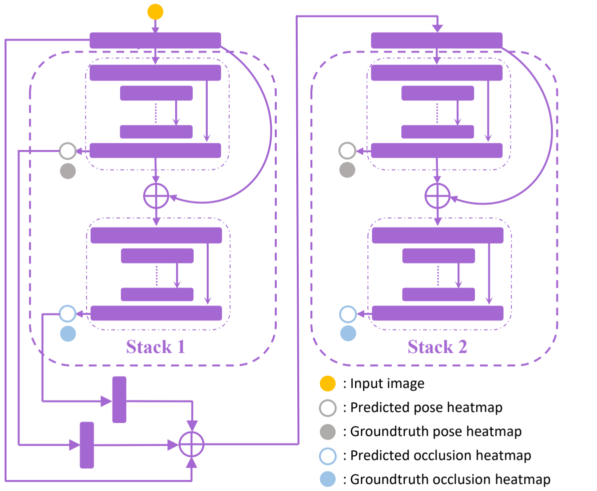
Stack 1
Stack 2
● : Input image
○ : Predicted pose heatmap
● : Groundtruth pose heatmap
○ : Predicted occlusion heatmap
● : Groundtruth occlusion heatmap

Figure 3: Architecture of the multi-task generative network *G*. Each rectangular block indicates a convolutional block. ⊕ indicates addition of input features. The stacking of the first and the second networks is shown and more networks can be stacked with the same structure.

In each basic block, the final heatmap outputs $\hat{\boldsymbol{y}}_n, \hat{\boldsymbol{z}}_n$ are obtained from $\boldsymbol{Y}_n$ and $\boldsymbol{Z}_n$ by two 1×1 convolution layers with the step size of 1 and without padding. Specifically, the first convolution layer reduces the number of feature maps from the number of feature maps to the number of body parts. The second convolution layer acts as a linear classifier to obtain the final predicted heatmaps.

Therefore, given a training set $\{\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{z}^i\}_{i=1}^M$ where M is the number of training images, the loss function of our multi-task generative network is presented as:

$$\mathcal{L}_G(\Theta) = \frac{1}{2MN} \sum_{n=1}^{N} \sum_{i=1}^{M} \left(\left\| \boldsymbol{y}^i - \hat{\boldsymbol{y}}_n^i \right\|^2 + \left\| \boldsymbol{z}^i - \hat{\boldsymbol{z}}_n^i \right\|^2 \right) . \tag{1}$$

where Θ denotes the parameter set.

2.2. Pose Discriminator

To enable the training of the network to exploit priors about the human body joints configurations, we design the pose discriminator *P*. The role of the discriminator *P* is to distinguish the *fake* poses (poses do not satisfy the constraints of human body joints) from the *real* poses.

It is intuitive that we need local image regions to identify the body parts and the large image patches (or the whole image) to understand the relationships between body parts. However, when some parts are seriously occluded, it can be very difficult to locate the body parts. Human can achieve that by using prior knowledge and observing both the local image patches around the body parts and relationships among different body parts. Inspired by this, both low-level and high-level information can be important to infer whether the predicted poses are biologically plausible. In contrast to previous work, we use an encoder-decoder architecture to implement the discriminator *P*. Skip connections between parallel layers are used to incorporate both the local and global information.

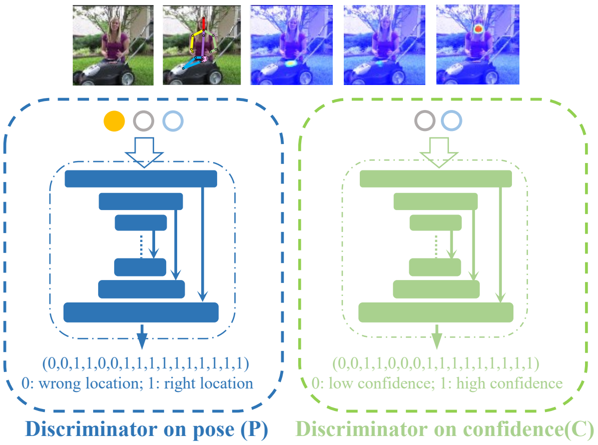
(0,0,1,1,0,0,1,1,1,1,1,1,1,1,1,1)
0: wrong location; 1: right location
(0,0,1,1,0,0,0,1,1,1,1,1,1,1,1,1)
0: low confidence; 1: high confidence
Discriminator on pose (P)
Discriminator on confidence(C)

Figure 4: Architectures of the discriminator network *P* and *C*. On the top we show the image for pose estimation, the image with estimated joints and heatmaps of right ankle, pelvis and neck (1st, 7th and 9th of all pose heatmaps respectively). The expected output for this sample is given in the bottom.

Additionally, even when the generative network fails to predict the correct pose locations for a particular image, the predicted pose may still be a plausible one for another human body shape. Thus, simply using the pose and occlusion features may still face difficulty in training an accurate *P*. *Such inference should be made by taking the original image into consideration at the same time.* Occlusion information can also be useful in inferring the pose rationality. So we use the input RGB image with pose and occlusion heatmaps generated by the *G* net as the input to *P* for predicting whether a pose is reasonable or not for a particular image. The network structure of *P* is shown in Fig. 4. To achieve this goal, GAN is set in the conditional manner for *P* in our framework. As GANs learn a generative model of data, conditional GANs (cGANs) learn a conditional generative model [12]. The objective of a conditional adversarial *P* network is expressed as follows:

$$\begin{aligned} \mathcal{L}_P(G, P) = \ & \mathbb{E}[\log P(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x})] + \\ & \mathbb{E}[\log(1 - |P(G(\boldsymbol{x}), \boldsymbol{x}) - \boldsymbol{p}_{\text{fake}}|)] . \end{aligned} \tag{2}$$

where $\boldsymbol{p}_{\text{fake}}$ is the ground truth pose discriminator label. In traditional GAN, $\boldsymbol{p}_{\text{fake}}$ is simply set as 0. The illustration of $\boldsymbol{p}_{\text{fake}}$ here will be discussed in detail in Section 2.4.

2.3. Confidence Discriminator

By observing the differences between ground truth heatmaps and predicted heatmaps by previous methods, we find that the predicted ones are often not Gaussian centered because of occlusions and body overlapping. Recalling the mechanism of human vision, even when the body parts are occluded, we can still confidently locate the body parts. That is mainly because we already acquire the geometric prior

Algorithm 1 The training process of our method.

```
Require: Training images: $\boldsymbol{x}$, the corresponding ground-truth heatmaps $\{\boldsymbol{y},\boldsymbol{z}\}$;
1: Forward P by $\{\hat{\boldsymbol{p}}_{\text{fake}}\} = P(\boldsymbol{x}, G(\boldsymbol{x}))$, and optimize P net by maximizing the second term in Eq. (2);
2: Forward P by $\{\hat{\boldsymbol{p}}_{\text{real}}\} = P(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, and optimize P by maximizing the first term in Eq. (2);
3: Forward C by $\{\hat{\boldsymbol{c}}_{\text{fake}}\} = C(G(\boldsymbol{x}))$, and optimize C by maximizing the second term in Eq. (3);
4: Forward C by $\{\hat{\boldsymbol{c}}_{\text{real}}\} = C(\boldsymbol{y}, \boldsymbol{z})$, and optimize C by maximizing the first term in Eq. (3);
5: Optimize G by Eq. (4);
6: Go back to Step 1 until the accuracy of the validation set stop increasing;
7: return $G$.
```

of human body joints. Motivated by this, we design a second auxiliary discriminator, which is termed Confidence Discriminator (*i.e.*, *C*) to discriminate the high-confidence predictions from the low-confidence predictions. The inputs for *C* are the pose and occlusion heatmaps. The objective of a traditional adversarial *C* network can be expressed as:

$$\begin{aligned}\mathcal{L}_C(G, C) = \mathbb{E}[\log C(\boldsymbol{y}, \boldsymbol{z})] + \\ \mathbb{E}[\log(1 - |C(G(\boldsymbol{x})) - \boldsymbol{c}_{\text{fake}}|)]\,. \qquad (3)\end{aligned}$$

where $\boldsymbol{c}_{\text{fake}}$ is the ground truth confidence label. In traditional GAN, $\boldsymbol{c}_{\text{fake}}$ is simply set as 0. The illustration of $\boldsymbol{c}_{\text{fake}}$ here will also be discussed in Section 2.4.

2.4. Training of the Adversarial Networks

In this section, we describe in detail how *P* and *C* contribute to the accurate pose predictions with structure constraints.

First we show how to embed the geometric information of human bodies into the proposed *P* network. We observe that, when a part of human body is occluded, the prediction of the un-occluded parts are typically not affected. This may be due to the DCNN's strong ability in learning local features.

However, in previous works on image translation using GANs, the discriminative network is learned with all fake samples being labeled 0. When predicted heatmaps are close enough to groundtruths, considering it as a successful prediction makes sense. We also found the network to be difficult to converge by simply setting 0 or 1 as ground truth for a sample. Based on these observations, we designed a novel strategy for pose estimation. This leads to the difference with traditional GANs as in Eq. (2) and Eq. (3).

The ground truth $\boldsymbol{p}_{\text{real}}$ of a real sample is a 16×1 unit vector. For the fake samples, if a predicted body part is far from the ground truth location, the pose is clearly implausible for the body configuration in this image. Therefore, when training P, the ground truth $\boldsymbol{p}_{\text{fake}}$ is:

$$\boldsymbol{p}^i_{\text{fake}} = \begin{cases} 1 & \text{if } d_i < \delta \\ 0 & \text{if } d_i \geqslant \delta \end{cases},$$

where δ is the threshold parameter and d_i is the normalized distance between the predicted and ground-truth location of the i-th body part. The range of the output values in *P* is also $[0, 1]$. To deceive *P*, *G* will be trained to generate heatmaps which satisfy the joints constraints of human bodies.

As mentioned in Section 2.2 and Section 2.3, the previous pose estimation networks usually have less confidences in locating the occluded body parts as the local information are neglected. However, if the *G* network can learn to make inferences like human in this situation, it should achieve higher confidences in locating such body parts.

If *G* generates low-confidence heatmaps, *C* will classify the result as "fake". As *G* is optimized to deceive *C* that the fakes being real, this process would help *G* to generate high confidence heatmaps even with occlusions presented. The outputs are the confidence scores $\boldsymbol{c}$ which in fact corresponds to whether the network is confident in locating body parts.

During training *C*, the real heatmaps are labelled with a 16×1 (16 is the number of body parts) unit vector $\boldsymbol{c}_{\text{real}}$. The confidence of the fake (predicted) heatmap should be high when it is close to ground truth and low otherwise, instead of being low for all predicted heatmaps as in traditional GANs. So the fake (predicted) heatmaps are labelled with a 16×1 vector $\boldsymbol{c}_{\text{fake}}$ where the elements of $\boldsymbol{c}_{\text{fake}}$ are the corresponding confidence scores.

$$\boldsymbol{c}^i_{\text{fake}} = \begin{cases} 1 & \text{if } \|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\| < \varepsilon \\ 0 & \text{if } \|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\| \geqslant \varepsilon \end{cases},$$

where ε is the threshold parameter, and i is the i-th body part. The range of the output values in *C* is $[0, 1]$.

Previous approaches to conditional GANs have found it beneficial to mix the GAN objective with a traditional loss, such as ℓ_2 distance [21]. For our task, it is clear that we also need to supervise *G* in the training process with the ground truth human poses. Thus, the discriminator still plays the original role, but the generator will not only fool the discriminator but also approximate the ground-truth output in an ℓ_2 sense as in Eq. (3). Therefore, the final objective function is presented as follows.

$$\arg\min_{G}\max_{P,C} \mathcal{L}_G(\Theta) + \alpha\mathcal{L}_C(G, C) + \beta\mathcal{L}_P(G, P)\,. \qquad (4)$$

$\alpha = 0$ if $\boldsymbol{c}_{\text{fake}} = \boldsymbol{c}_{\text{real}}$, $\beta = 0$ if $\boldsymbol{p}_{\text{fake}} = \boldsymbol{p}_{\text{real}}$. In experiments, in order to make the different components of the final objective function have the same scale, the hyper parameters α and β are set to $1/220$ and $1/180$, respectively. Algorithm 1 demonstrates the whole training processing as the pseudo codes.

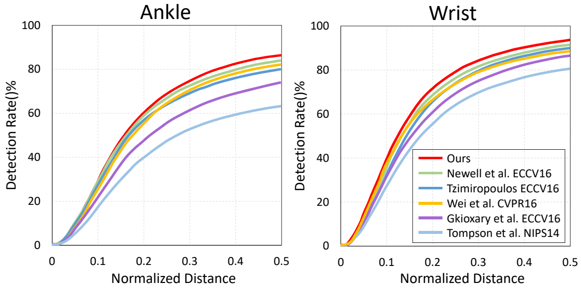

Figure 5: PCKh comparison on MPII validation set.

3. Experiments

Datasets. We evaluate the proposed method on two widely used benchmarks on pose estimation, *i.e.*, extended Leeds Sports Poses (LSP) [16] and MPII Human Pose [1]. The LSP dataset consists of 11k training images and 1k testing images from sports activities. The MPII dataset consists of around 25k images with 40k annotated samples (about 28k for training, 11k for testing). The figures are annotated with 16 landmarks on the whole body with various challenging directions to the camera. On MPII, we train our model on a subset of training images while evaluating on the official test set and a held-out validation set about 3000 samples [30, 20]. Both datasets provide the visibility of body parts, which is used as the supervision occlusion signal in our method.

Experimental Settings. According to the rough person location given by the dataset, we crop the images with the target human centered at the images, and warp the image patch to the size of 256×256 pixels. We follow the data augmentation in [20] by rotation (+/- 30 degrees), and scaling (0.75-1.25). During training for LSP, we use the MPII dataset to augment the training data of LSP, which is a regular routine as done in [34, 15].

During testing on the MPII dataset, we follow the standard routine to crop image patches with the given rough position and scale. The network starts with a 7×7 convolutional layer with stride 2, followed by a residual modules and a max pooling to drop the resolution down from 256 to 64. Then two residual modules are followed before sending the feature into G. Across the entire network all residual modules contain three convolution layers and a skip connection with output of 512 feature maps. The generator is stacked four times if not specially indicated in our experiment. For implementation, we train our model with the Torch7 toolbox [8]. The network is trained using the RMSprop algorithm with initial learning rate of 2.5×10^{-4}. The model on the MPII dataset was trained for 230 epochs and the LSP dataset for 250 epochs (about 1 and 1.5 days on a Tesla M40 GPU).

3.1. Quantitative Results

We use the Percentage Correct Keypoints (PCK@0.2) [37] metric for comparison on the LSP dataset which reports the percentage of detection that falls within a normalized distance of the ground-truth for comparisons. For MPII, the distance is normalized by a fraction of the head size [1] (referred to as PCKh@0.5).

Table 1: Comparisons of PCK@0.2 performance on the LSP dataset.

Methods	*Head*	*Sho.*	*Elb.*	*Wri.*	*Hip*	*Knee*	*Ank.*	**Mean**
[2]	95.2	89.0	81.5	77.0	83.7	87.0	82.8	85.2
[17]	96.8	89.0	82.7	79.1	90.9	86.0	82.5	86.7
[22]	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1
[15]	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
[23]	97.8	92.5	87.0	83.9	91.5	89.9	87.2	90.1
[34]	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
[4]	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Ours	**98.5**	**94.0**	**89.8**	**87.5**	**93.9**	**94.1**	**93.0**	**93.1**

Table 2: Results on MPII Human Pose (PCKh@0.5).

Methods	*Head*	*Sho.*	*Elb.*	*Wri.*	*Hip*	*Knee*	*Ank.*	**Mean**
[31]	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
[5]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
[30]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
[14]	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
[22]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
[17]	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
[11]	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
[25]	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
[15]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
[34]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
[4]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
[20]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
[7]	**98.5**	96.3	91.9	88.1	**90.6**	88.0	85.0	91.5
Ours(test) [1]	98.1	**96.5**	**92.5**	**88.5**	90.2	**89.6**	**86.0**	**91.9**
Ours(-valid)[2]	98.2	96.2	90.9	86.7	89.8	87.0	83.2	90.6
Ours(valid)[3]	98.6	96.4	92.4	88.6	91.5	88.6	85.7	92.1

[1] Our full model on test set [2] Our baseline model on validation set
[3] Our full model on validation set

LSP Human Pose. Table 1 shows the PCK performance of our method and previous methods at a normalized distance of 0.2. Our approach outperforms the state-of-the-art across all the body joints, and obtains 2.4% improvement in average.

MPII Human Pose. Table 2 and Fig. 5 reports the PCKh performance of our method and previous methods at a normalized distance of 0.5. Baseline model refers to a four-stacked single-task network without multi-task and discriminators. It has similar structure but half of stacked layers and parameter numbers compared to [20]. Our method achieves the best PCKh score of 91.9% on the test set.

In particular, for the most challenging body parts, *e.g.*, wrist and ankle, our method achieves 0.4% and 1.0% improvement compared with the closest competitor respectively.

3.2. Quanlitative Comparisons

To gain insights on how the proposed method accomplish the goal of setting the pose estimations within the geometric constraints, we visualize the predicted poses on the MPII test set compared with a 2-stacked hourglass network (HG) [20],

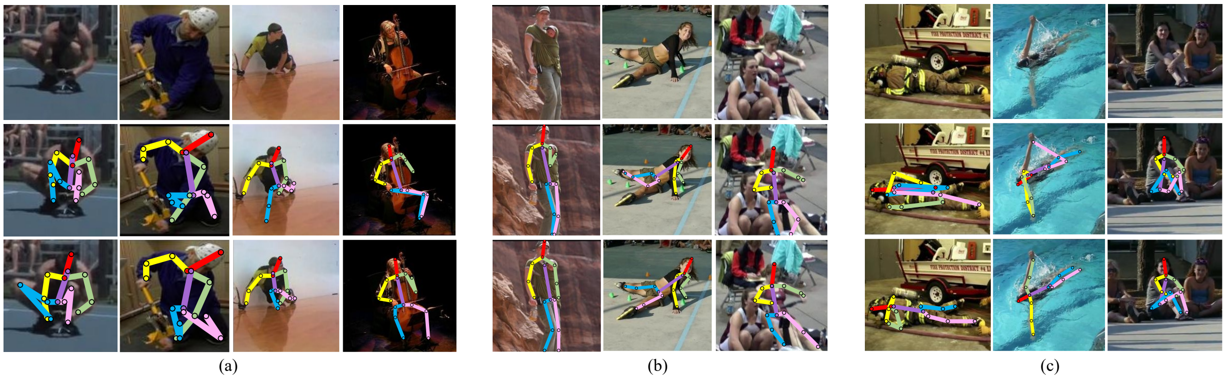

Figure 6: Prediction samples on the MPII test set. The first row: original images. The second row: results by stacked hourglass network (HG) [20]. The third row: results by our method. (a)-(c) stand for three kinds of failure with HG.

as demonstrated in Fig. 6. For fair comparison, we also use a 2-stacked network in this section. We can see that our method gains a better understanding of the human body which leads to less weird predictions.

In (a), the human body is highly twisted or partly occluded, which results in some invisible body limbs. In these cases, HG fails to understand some poses while our method succeeds. This may be because of the ability of occlusion prediction and shape prior learned the in the training process. In (b), HG locates some body parts to the nearby positions with the most salient features. This indicates that HG has learned excellent features about body parts. However, without human body structure awareness, this may locate some body parts to the surrounding area instead of the right one. In (c), due to lack of body configuration constraints, HG produces poses with weird twisting across body limbs. As we have implicitly embedded the body constraints into our discriminator, our network succeeds in predicting the correct body location even under some difficult situations.

On the other hand, we also show some failure examples of our method on the MPII test set in Fig. 7. As shown in Fig. 7, our method may fail in some challenging cases with twisted limbs at the edge, overlapping people and occluded body parts. In some cases, human may also fail to figure out the correct pose at a glance. Even when our method fails in this situations, it also achieves more reasonable poses compared to previous method. Previous method may generate some poses which violate human body structure as shown in the first row of Fig. 7. When the network fails to find high-confidence locations around the person, it shifts to the surrounding area where the local features matches the trained features best. Lacking of shape constraint finally results in these absurd poses.

3.3. Occlusion Analysis

Here we present a detailed analysis of the outputs of the networks when joints in the images are occluded.

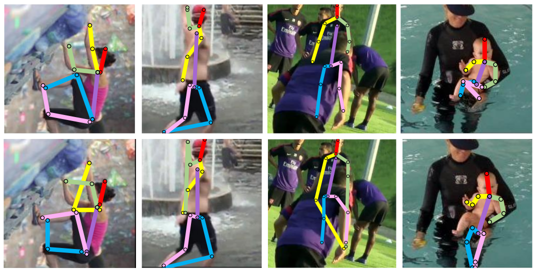

Figure 7: Failure cases caused by body part at the edge (the first-second columns), overlapping people (the third column) and invisible limbs (the fourth column). The results on the first and second rows are generated by our method and HG [20], respectively.

First, two examples with some body parts occluded are given in Fig. 8. In the first sample, two legs of the person are totally occluded by the table. In the corresponding occlusion maps, the occluded part are well predicted. Despite of the occlusions, the pose heatmaps generated by our method are mostly clear and Gaussian centered. This results in high scores in both pose prediction and confidence evaluation despite of occlusions.

In the second image, half part of the person is overlapped by the person ahead of him. Our method also succeeds to yield the correct pose locations with clear heatmaps. Occlusion information is also well predicted for the occluded parts. As shown in the columns in red, although the confidence scores of the occluded body parts are comparatively low, they remain an overall high level. This shows that our network has learned some human body priors during training. Thus it has the ability to predict reasonable poses even under some occlusions. This verifies our motivation of designing the discriminators with GANs.

Next, we compare the performance of our method under occlusions with a stacked hourglass network [20] as the strong baseline. In the validation set of MPII, about 25% of the elbows and wrists with annotations are labeled invisible. We show the results of elbows and wrists with visible samples and invisible samples in Table 3. For body parts without occlusions, our method improves the baseline by about 0.8% of detection rate. However, *our method improves the baseline by 3.5% and 3.6% of detection rates on the*

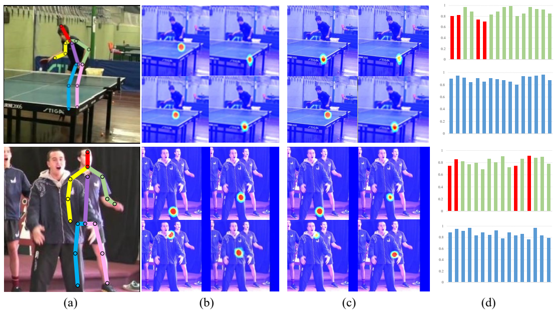

Figure 8: (a) Input images with predicted poses; (b) Predicted pose heatmaps of four occluded body parts; (c) Predicted occlusion heatmaps of four occluded body parts; (d) Outputs values of P (in blue) and C(in green). Red columns in the output of C correspond to values of the four occluded body parts.

invisible wrists and elbows. This shows the advantage of our method in dealing with body parts with occlusions.

3.4. Ablation Study

To investigate the efficacy of the proposed multi-task generator network and the discriminators designed for learning human body priors, we conduct ablation experiments on the validation set of the MPII Human Pose dataset. A four-stacked single-task generator without occlusion is used as the baseline. The overall result is shown in Fig. 9. We give analysis to two components in our method: the multi-task manner and discriminators.

Multi-task. We compare the four-stacked multi-task generator with the baseline. The networks are trained by removing the discriminators (*i.e.*, no GANs). By using the occlusion information, the performance on the MPII validation set increases 0.5% compared to the baseline model. This shows that the multi-task structure helps the network to understand the poses.

Discriminator with Single-task. We also compare the four-stacked single-task generator trained with discriminators with the baseline. The networks are trained by removing the part for the occlusion heatmaps. Discriminators also receive inputs without occlusion heatmaps. By using the body-structure-aware GANs, the performance on the MPII validation set increases by 0.6% compared to the baseline model. This shows that the discriminators contribute in pushing the generator to produce more reliable pose predictions.

In general, individually adding the multi-task or discriminator both increase the accuracy of location. But using them separately results in 0.6% and 0.5% improvement respectively, while using both produces an improvement of 1.5%. The reliability of P and C on sufficient feature to discriminate the results may be the reason. Occlusion features obviously can help to understand the image and the generated pose for the discriminators.

Table 3: Detection rates of visible and invisible elbows and wrists.

Methods	Visible		Invisible	
	Wrist	*Elbow*	*Wrist*	*Elbow*
[20]	93.6	95.1	67.2	74.0
Ours	**94.5**	**95.9**	**70.7**	**77.6**

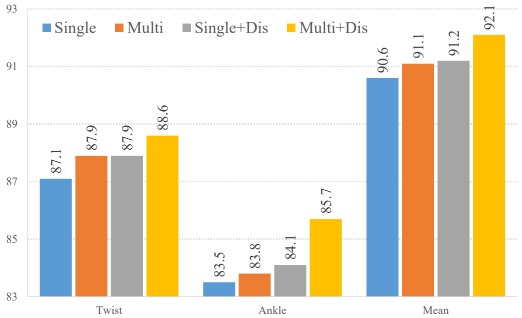

Figure 9: Ablation study: PCKh scores at the threshold of 0.5.

4. Conclusions

In this paper, we proposed a novel conditional adversarial network for pose estimation, termed Adversarial PoseNet, which trains a multi-task pose generator with two discriminator networks. The two discriminators function like an expert who distinguishes reasonable poses from unreasonable ones. By training the multi-task pose generator to deceive the expert that the generated pose is real, our network is more robust to occlusions, overlapping and twisting of human bodies. In contrast to previous work using DCNNs in human pose estimation, our network is able to alleviate the risk of locating the human body part onto the matched features without consideration of human body priors.

Although we need to train three sub-networks (G, P, C), we only need to use G net during testing. With a small computation overhead, we achieve considerably better results on two popular benchmark datasets. We have also verified that our network can produce human poses which are mostly within the manifold of human body shape.

The method developed here can be immediately applied to other shape estimation problems such as face landmark detection using DCNNs. More significantly, we believe that the use of GANs as a tool to predict structured output or enforcing output dependency can be further developed to much more general structured output learning.

Acknowledgements

Y. Chen and X.-S. Wei's participation was in part support by China Scholarship Council. C. Shen's participation was in part supported by an ARC Future Fellowship. L. Liu's participation was in part supported by an ARC DECRA Fellowship. J. Yang's participation was supported by the National Science Fund of China under Grants #91420201, 61472187, 61502235, 61233011, 61373063 and 61602244, the 973 Program #2014CB349303, Program for Changjiang Scholars and Innovative Research Team in University.

References

[1] M. Andriluka, L. Pishchulin, P. V. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3686–3693, 2014.

[2] V. Belagiannis and A. Zisserman. Recurrent human pose estimation. In *Proc. IEEE Int. Automatic Face & Gesture Recognition*, 2017.

[3] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Upper body detection and tracking in extended signing sequences. *Int. J. Comput. Vision*, 95(2):180–197, 2011.

[4] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *Proc. Eur. Conf. Comp. Vis.*, pages 717–732, 2016.

[5] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4733–4742, 2016.

[6] X. Chu, W. Ouyang, H. Li, and X. Wang. Structured feature learning for pose estimation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4715–4723, 2016.

[7] X. Chu, W. Ouyang, H. Li, and X. Wang. Multi-context attention for human pose estimation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.

[8] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn Workshop Advances in Neural Inf. Process. Syst.*, 2011.

[9] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 1486–1494, 2015.

[10] M. Eichner, M. J. Marín-Jiménez, A. Zisserman, and V. Ferrari. 2D articulated human pose estimation and retrieval in (almost) unconstrained still images. *Int. J. Comput. Vision*, 99(2):190–214, 2012.

[11] G. Gkioxari, A. Toshev, and N. Jaitly. Chained predictions using convolutional neural networks. In *Proc. Eur. Conf. Comp. Vis.*, pages 728–743, 2016.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 2672–2680, 2014.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 770–778, 2016.

[14] P. Hu and D. Ramanan. Bottom-up and top-down reasoning with hierarchical rectified Gaussians. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5600–5609, 2016.

[15] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. DeeperCut: A deeper, stronger, and faster multi-person pose estimation model. In *Proc. Eur. Conf. Comp. Vis.*, pages 34–50, 2016.

[16] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proc. British Machine Vision Conf.*, 2010. doi:10.5244/C.24.12.

[17] I. Lifshitz, E. Fetaya, and S. Ullman. Human pose estimation using deep consensus voting. In *Proc. Eur. Conf. Comp. Vis.*, pages 246–260, 2016.

[18] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. In *Proc. Int. Conf. Learn. Representations*, pages 1–14, 2016.

[19] M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. *arXiv preprint arXiv*, 1411.1784:1–7, 2014.

[20] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *Proc. Eur. Conf. Comp. Vis.*, pages 483–499, 2016.

[21] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2536–2544, 2016.

[22] L. Pishchulin, M. Andriluka, P. V. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 3487–3494, 2013.

[23] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. DeepCut: joint subset partition and labeling for multi person pose estimation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4929–4937, 2016.

[24] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv*, 1511.06434:1–16, 2015.

[25] U. Rafi, I. Kostrikov, J. Gall, and B. Leibe. An efficient convolutional network for human pose estimation. In *Proc. British Machine Vis. Conf.*, pages 1–11, 2016.

[26] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *Proc. Int. Conf. Mach. Learn.*, pages 1–10, 2016.

[27] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 2226–2234, 2016.

[28] B. Sapp and B. Taskar. MODEC: Multimodal decomposable models for human pose estimation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3674–3681, 2013.

[29] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *Proc. Eur. Conf. Comp. Vis.*, pages 406–420, 2010.

[30] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 648–656, 2015.

[31] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 1799–1807, 2014.

[32] A. Toshev and C. Szegedy. DeepPose: human pose estimation via deep neural networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1653–1660, 2014.

[33] X. Wang and A. Gupta. Generative image modeling using style and structure adversarial networks. In *Proc. Eur. Conf. Comp. Vis.*, pages 318–335, 2016.

[34] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4724–4732, 2016.

[35] W. Yang, W. Ouyang, H. Li, and X. Wang. End-to-end jearning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3073–3082, 2016.

[36] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1385–1392, 2011.

[37] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2878–2890, 2013.

[38] D. Yoo, N. Kim, S. Park, A. S. Paek, and I.-S. Kweon. Pixel-level domain transfer. In *Proc. Eur. Conf. Comp. Vis.*, pages 517–532, 2016.

[39] J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial network. In *Proc. Int. Conf. Learn. Representations*, 2017.

[40] Y. Zhou and T. L. Berg. Learning temporal transformations from time-lapse videos. In *Proc. Eur. Conf. Comp. Vis.*, pages 262–277, 2016.