ACCEPTED VERSION

Yu Chen, Chunhua Shen, Hao Chen, Xiu-Shen Wei, Lingqiao Liu, Jian Yang

Adversarial learning of structure-aware fully convolutional networks for landmark localization

Published version at: http://dx.doi.org/10.1109/TPAMI.2019.2901875

http://hdl.handle.net/2440/126991

Adversarial Learning of Structure-Aware Fully Convolutional Networks for Landmark Localization

Yu Chen[1], Chunhua Shen[2], Hao Chen[2], Xiu-Shen Wei[3], Lingqiao Liu[2] and Jian Yang[1]

Abstract—Landmark/pose estimation in single monocular images has received much effort in computer vision due to its important applications. It remains a challenging task when input images come with severe occlusions caused by, e.g., adverse camera views. Under such circumstances, biologically implausible pose predictions may be produced. In contrast, human vision is able to predict poses by exploiting geometric constraints of landmark point inter-connectivity. To address the problem, by incorporating priors about the structure of pose components, we propose a novel structure-aware fully convolutional network to implicitly take such priors into account during training of the deep network. Explicit learning of such constraints is typically challenging. Instead, inspired by how human identifies implausible poses, 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. Training of the network follows the strategy of conditional Generative Adversarial Networks (GANs). The effectiveness of the proposed network is evaluated on three pose-related tasks: 2D human pose estimation, 2D facial landmark estimation and 3D human pose estimation. The proposed approach significantly outperforms several state-of-the-art methods and almost always generates plausible pose predictions, demonstrating the usefulness of implicit learning of structures using GANs.

Index Terms—Pose Estimation, Landmark Localization, Structure-aware Network, Adversarial Training, Multi-task Learning

✦

1 INTRODUCTION

Landmark localization, a.k.a, keypoint localization, pose estimation or alignment (we use these terms interchangeably in the sequel), is a key step in many vision tasks. For example, face alignment, which is to locate the positions of a set of predefined facial landmarks from a single monocular facial image, plays an important role for facial augmented reality and face recognition. Human pose prediction locates the positions of a few human body joints, which is critically important in understanding the actions and emotions of people in images and videos. Keypoint prediction from monocular images is a challenging task due to factors such as high flexibility of facial/body limbs deformation, self and outer occlusion, various camera angles, etc. In this work, we consider the problem of human pose estimation and facial landmark detection in the same framework with minimum modification as essentially they both are image-to-point regression problems. We achieved state-of-the-art on both tasks at the submission of this manuscript.

Recently, significant improvements have been achieved on 2D pose estimation by using Deep Convolutional Neural Networks (DCNNs) [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. These approaches mainly follow the strategy of regressing heatmaps or landmark coordinates of each pose part using DCNNs. These regression models have shown great ability in learning better feature representations. However, for pose components with heavy occlusions and background clutters that appear similar to body parts, DCNNs may have difficulty in regressing accurate poses.

Human vision is capable of learning the shape structures from abundant observations. Even under extreme occlusions, one can infer the potential poses and exclude the implausible ones. It is, however, very challenging to incorporate the priors about shape structures into DCNNs, because, as pointed out in [4], the low-level mechanics of DCNNs is typically difficult to interpret, and DCNNs are most capable of learning features.

As a consequence, an unreasonable pose may be produced by conventional DCNNs. As shown in Fig. 1, in challenging test cases with heavy occlusions, standard DCNNs tend to perform poorly. To tackle this problem, priors about the structure of the body joints should be taken into account. The key to this problem is to learn the *real* body joints distribution from a large amount of training data. However, explicit learning of such a distribution is not trivial.

To address this problem, we attempt to learn the distribution of the human body structures *implicitly*. Similar to the human vision, we suppose that we have a "discriminator" which can tell whether the predicted pose is geometrically plausible. 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) [11], [12], [13], [14], [15], we propose to

• [1] *Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology;* [2] *School of Computer Science, The University of Adelaide, Australia;* [3] *Department of Computer Science and Technology, Nanjing University, China.*

• *This work was done when Y. Chen and X-S. Wei were visiting The University of Adelaide. Correspondence should be addressed to C. Shen and J. Yang.*
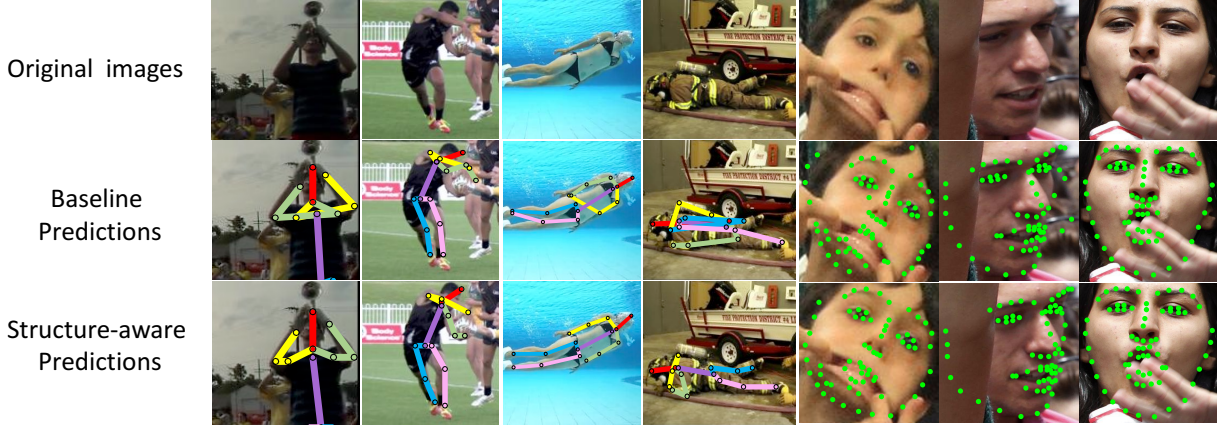
Figure 1. **Motivation**. We show the importance of strongly enforcing priors about the pose structure during training of DCNNs for pose estimation. Learning without using such priors generates inaccurate results.

design the "discriminator" as the discriminator network in GAN while the regression network functions as the generative network. Training the generator in an adversarial manner against the discriminator precisely meets our intention.

For both 2D human pose estimation and facial landmark localization, a baseline stacked bottom-up, top-down networks G is designed to generate the pose heatmaps. Based on the pose heatmaps, the pose discriminator (P) is used to tell whether the pose configuration is plausible. The generator is asked to "fool" the discriminators by training G and P in the generative adversarial manner. Thus, the human body structure is implied in the P net by guiding G to the direction that is close to ground-truth heatmaps and satisfies 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 is required.

What is more, the function of the discriminator is not limited to heatmap regression based 2D pose estimation. For tasks concerning structured outputs (*e.g.*, 2D to 3D human pose transformation), we can easily extend our method by using the adversarial discriminator on a baseline method to learn the 3D structure distributions for generating more plausible 3D pose prediction, as we show in our experiments.

The main contributions of this work are thus as follows

- To our knowledge, we are the first to use Generative Adversarial Networks (GANs) to exploit the constrained pose distribution for improving pose estimation. We also design a stacked multi-task network for predicting both the pose heatmaps and the occlusion heatmaps to achieve improved results for 2D human pose estimation.
- We design a novel network framework for pose estimation which takes the geometric constraints of keypoints 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.

- We evaluate our method on public 2D human pose estimation datasets, 2D facial landmark estimation datasets and 3D human pose estimation datasets. Our approach achieved state-of-the-art performance at the submission of this manuscript, and is able to consistently produce more plausible pose predictions compared to baseline methods.

Furthermore, concurrently with recent work of [16], we may be one of the first to directly use DCNNs to regress heatmaps for facial landmark estimation. Due to the help of the structure-aware network structure, the traditional complex cascaded procedure is avoided.

2 RELATED WORK

The task of human pose estimation can be divided into multi-person and single-person. Multi-person pose estimation involves both human detection and pose estimation. The difficulty lies in accurate detection of individuals with different poses, overlapping or occlusions. While in single-person human pose estimation, the rough positions of the person can be easily obtained. The main challenge in single-person pose estimation is pose variation caused by body motion, etc. As our method focuses on the positive influence of adversarial learning by exploiting the structure of pose estimation, we only consider single-person pose estimations and multi-person pose estimations with given person-detection results in this work. In terms of vision tasks, our method is mostly related to 2D and 3D human pose estimation and 2D facial landmark estimation issues. In terms of mechanism of deep learning models, our method is mostly related to the Generative Adversarial Networks.

2D Human Pose Estimation. Traditional 2D single human pose estimation methods often follow the framework of tree structured graphical model [17], [18], [19], [20], [21], [22]. With the introduction of "DeepPose" by Toshev *et al*. [6], deep network based methods become more popular in this area. This work is closely related to the methods generating pose heatmaps from images [4], [5], [7], [8], [9], [23], [24], [25]. For example, Tompson *et al*. [4] used multi-resolution feature representations to generate heatmaps with joint-training of a Markov Random Field (MRF). Tompson
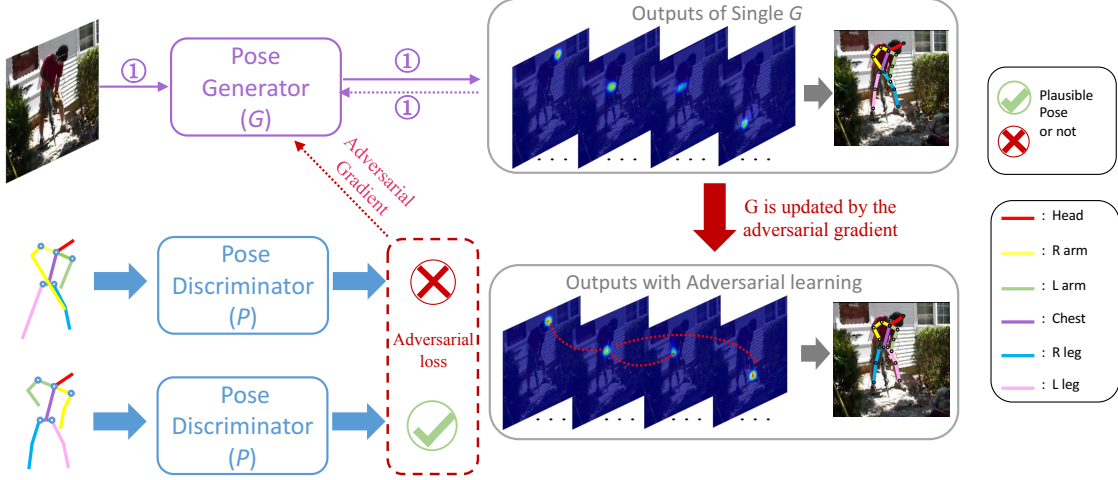
Figure 2. Overview of the proposed Structure-aware Convolutional Network for human pose estimation. The sub-network in purple is the stacked multi-task network (*G*) for pose generation. The networks in blue (*P*) is used to discriminate whether the generated pose is "real" (reasonable as a body shape). The loss of *G* has two parts: Mean Squared Error (MSE) of heatmaps (dashed line in purple) and Binary Cross Entropy (BCE) adversarial loss from *P* (dashed line in red). Standalone training of *G* produces results in the top-right. *G* and *P* produce results at the bottom-right.

et al. [5] used multiple branches of convolutional networks to fuse the features from an image pyramid, and used MRF for post-processing. Later, Convolutional Pose Machine [8] incorporated the inference of the spatial correlations among body parts within convolutional networks. The hourglass network [9] introduced a state-of-the-art architecture for bottom-up and top-down inference built upon residual blocks and skip connections. Based on the hourglass structure, Chu *et al.* [26] used convolutional neural networks with a multi-context attention mechanism in an end-to-end framework. The structure of our *G* net for this task is also a fully convolutional network with "conv-deconv" architecture as in [9]. However, our network is designed in a multi-task manner for improved performance.

Multi-person pose estimation methods mainly follow "bottom-up" or "top-down" architectures. The common "top-down" architecture [27], [28], [29], [30], [31] is to use a person detector first and then employ single-person pose estimation methods. In comparison, "bottom-up" methods detect all joints first and then group them into different subjects. One of the most popular bottom-up methods is [32] which proposed Part Affinity Fields to model the connectivity between joints. It is the first real-time multi-person estimation method and achieves best performance on the MSCOCO-2016 keypoint challenge. More recently, some top-down methods [33], [34] employ strong object detectors [35] and carefully-designed single-person pose estimation networks, which significantly outperform previous methods. As our method does not involve any detection algorithm, we also follow the top-down approach with given detections to evaluate the effectiveness of adversarial learning.

3D Human Pose Estimation. Based on the 2D human pose predictions, inferring 3D joints is to match the spatial position of the depicted person from 2D to 3D. This can be traced back to the early work by Lee *et al.* [36]. As the literature of this problem is vast with approaches in a variety of settings [37], here we only review recent works which are most relevant to ours using deep networks in the sequel.

The first category is to infer 3D body configurations by estimating body angles from images [38], [39]. These approaches avoid estimating 3D joint positions directly, which offer the advantage of constraining the pose in a human-like structure and having lower dimensionality. Recently, some systems have explored the possibility of directly inferring 3D poses from images with end-to-end deep architectures [40]. Pavlakos *et al.* [41] introduced a deep convolutional neural network based on the stacked hourglass architecture [9], which maps 2D joint probability heatmaps to probability distributions in the 3D space. Moreno-Noguer [42] represented 2D and 3D poses with $N \times N$ distance matrices (DMs) and regresses 2D DMs to 3D DMs.

The DM regression approach as well as the the volumetric approach of Pavlakos *et al.* assumes that direct regression from 2D keypoints to 3D keypoints is difficult. However, Martinez *et al.* [43] showed that a simple fully-connected network can perform very well for this direct regression. As this network is of a simple structure and achieves high performance, we use it as our baseline model. We demonstrate that the idea of enforcing the adversarial training on the baseline model also works well for this 2D-to-3D problem.

More recently, Yang *et al.* [44] proposed an adversarial learning method. Their method is also built upon [43] and has used the same training pipeline as in our method. The difference is that their method has three inputs: pose heatmaps, depth-maps and geometric descriptor, which are jointly sent to the discriminator for determining if a pose is plausible. Their method is specially designed for 3D human pose estimation.

2D Face Landmark Estimation. Traditional regression based methods often follow a cascaded manner to update the landmark localization results in a coarse-to-fine fashion. This strategy has been proven to be very effective for face alignment. Early methods mainly use random forest regression as the regressors due to computational efficiency [45], [46], [47], [48]. Burgos-Artizzu et al. [49] proposed Robust Cascaded Pose Regression (RCPR) which improves robustness to outliers by detecting occlusions explicitly. Different from previous learning process, Supervised Descent Method (SDM)

[50] attempts to directly minimize the feature deviation between estimated and ground-truth landmarks which is finally induced into a simple linear regression problem with supervised descent direction. To accelerate the speed of SDM and overcome the drawbacks of handcrafted features, Local Binary Features (LBF) [51] are learned for linear regression by using the regression forest. Project-Out Cascaded Regression [52] was proposed by learning and employing a sequence of averaged Jacobians and descent directions in a subspace orthogonal to the facial appearance variation.

Recently, deep neural networks were also introduced for face alignment [53], [54], [55], [56]. These methods use deep networks to replace the traditional regressors but still follow the cascaded framework. It is worth pointing out that the Mnemonic Descent Method (MDM) [2] showed that for face alignment end-to-end training of a convolutional recurrent neural network architecture works well. The original cascaded steps are connected by recurrent connections and handcrafted features are replaced by convolutional features. We take a further step by directly regressing the landmark heatmaps from the face image. This approach of direct regressing was considered inefficient and unrealistic by most previous methods in the literature, as face shape is complex. However, we show that with the help of the adversarial learning, shape priors can be better captured, and good localization results are achieved.

Generative Adversarial Network. Generative Adversarial Networks have been widely studied in the literature for discrete labels [57], text [58] and also images. The conditional models have tackled inpainting [59], image prediction from a normal map [60], future frame prediction [61], future state prediction [62], product photo generation [63], and style transfer [64].

Human pose estimation can be considered as a translation from a RGB image to a multi-channel 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 constraints into the model. Thus we have implemented different training strategies for fake samples from traditional GANs. In the next section, we provide details.

3 ADVERSARIAL LEARNING FOR LANDMARK LOCALIZATION

As depicted in Fig. 2, the adversarial training model consists of two parts, *i.e.*, the pose generator network G and the pose discriminator network P. Without discriminators, G will be updated simply by backward propagation of itself (cf., the lines with ① in Fig. 2). This is defined as the *baseline model* for all the tasks. Thus, incorrect location pose estimations may be generated. It is necessary to leverage the power of discriminators to correct these incorrect estimations. Therefore, a discriminator network P is introduced into the framework.

After updating G by training with P in the adversarial manner (cf. the red dashed lines), the pose priors are implicitly exploited. *In practical training, the two parts of the loss are added together to optimize for G at the same time.*

Next, we first introduce the structure of the baseline generative networks and the discriminator networks. Then, we show the adversarial training paradigm.

3.1 Generative Network

In this section, we present the generative network G (baseline model) in our framework. For 2D human pose estimation and facial landmark localization, the networks are fully-convolutional which predict pose estimations from images in an end-to-end manner. For 3D human pose estimation, we use a fully-connected network for 2D-3D coordinate transformation based upon 2D predictions.

2D human pose estimation. To solve the problem of human pose estimation, it can be very beneficial to employ local evidence for identifying features for human joints. Meanwhile, it clear to see that a coherent understanding of the full body image must be in place to achieve good pose estimation. In addition, as reported in [8], large contextual regions are important for locating body parts. Hence the contextual region of a neuron, which is its receptive field, should be large. To achieve these goals, an "encoder-decoder" architecture is used. Also, to capture information at each scale, mirrored layers in the encoder and decoder are added, as shown in the bottom-right part of Fig. 3. Inspired by [9], our network can also be stacked to provide the network with the ability to re-evaluate the previous estimates and features. In each module of the G net, a residual block [65] is used for the convolution operator.

Besides, 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. As shown in Fig. 3, in each stacking module, poses and occlusions are jointly predicted. Then, both predictions are re-evaluated for the next stacking.

So 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. 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. The multi-task generative network can be formulated as:

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

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
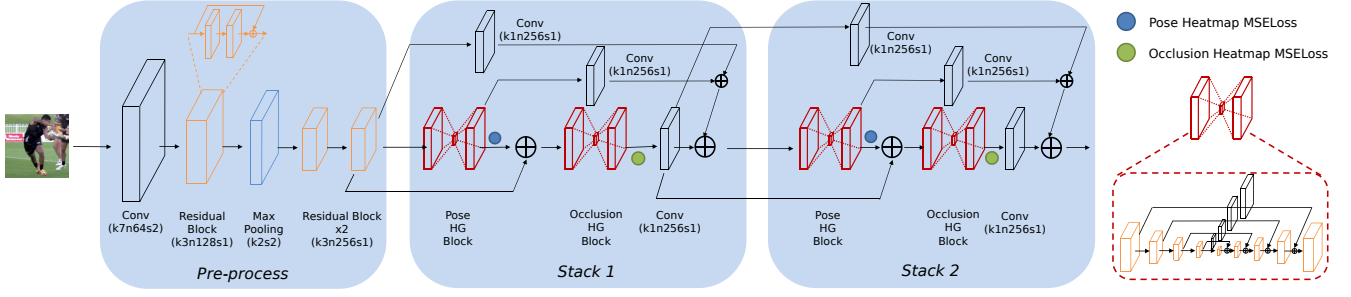
Figure 3. Architecture of the multi-task generative network *G*. Black, orange, blue and red rectangles indicate convolutional layers, residual blocks, max pooling layers and hourglass blocks respectively. ⊕ indicates addition of input features. Solid blue and green circles indicate pose and occlusion losses in the network. The brief architecture of the hourglass block is shown at the right. Stacking of the first and the second networks is displayed and more networks can be stacked with the same structure.

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:

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

where Θ denotes the parameter set.

2D Facial landmark localization. In contrast to most previous methods which predicts facial landmark location as coordinates, we use the same heatmap regression approach for human pose estimations. The variations of face shapes are clearly less complicated than human poses and most facial landmark databases do not contain visibility annotations. Therefore, we remove the occlusion heatmap regression part in Fig. 3 as the baseline for facial landmark localization. Thus, the network becomes a stacked hourglass architecture which is the same as in [9].

3D Human Pose Estimation. In this paper, 3D human poses are not predicted from scratch and are used as an extended validation of adversarial learning for 3D structure learning. To be specific, we follow [43] which deals 3D pose estimation as a 2-step procedure. First, heatmap based 2D predictions are given using fully-convolutional networks. Then 2D coordinates are extracted by extracting the locations of the maximum values in the heatmaps. Finally the 2D-3D coordinate transformation is done by combinations of linear layers followed by batch normalization, dropout and ReLU activation functions. To fully understand the network structure, readers may refer to [43] for details.

3.2 Discriminative Networks

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

2D Pose Discriminator. 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
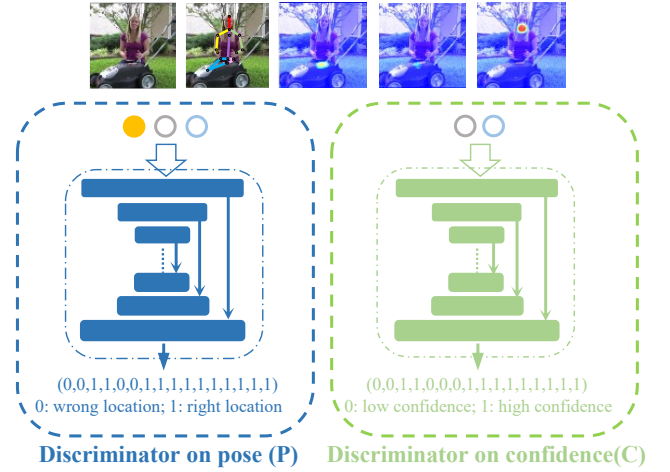
Figure 4. Architectures of the 2D pose discriminator networks *P* and *C*. On the top we show the image for pose estimation, the image with estimated joints and heatmaps of right ankle, pelvis and neck (1st, 7th and 9th of all pose heatmaps respectively). The expected output for this sample is given at the bottom of the dashed box.

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.

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.* When occlusion information can be provided, it is also helpful in inferring the pose rationality. Thus 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 for 2D human pose estimation. The network structure of *P* is shown in Fig. 4. To achieve this goal, GAN is designed in

Figure 5. Structure of the 3D pose discriminator when a 16-joints pose is to be discriminated. In each black rectangle, a few modules are combined.

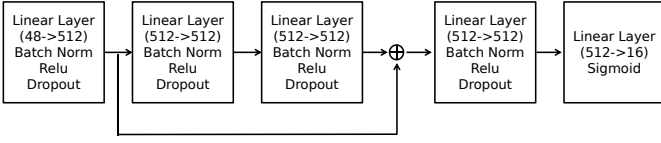the conditional manner for P in our framework. As GANs learn a generative model of data, conditional GANs (cGANs) learn a conditional generative model [14]. The objective of a conditional adversarial P network is expressed as follows:

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

where $\boldsymbol{p}_{\text{fake}}$ is the ground truth pose discriminator label. In traditional GAN, $\boldsymbol{p}_{\text{fake}}$ is simply set to 0. Detailed discussions of $\boldsymbol{p}_{\text{fake}}$ are presented in Section 3.3.

3D Pose Discriminator. Different from 2D pose estimations, 3D poses are in the form of 3-dimensional coordinates. In consistency with the simple structure in 2D-3D transformation, a five-layer fully connected network is used as the 3D Pose Discriminator as shown in Fig. 5. The output of this discriminator is the same as the one for 2D and also follows the same objective functions as in Eq. (2).

Auxiliary 2D Confidence Discriminator for 2D human pose. For the task of human pose estimation, some body parts are out of the boundary of the images and the corresponding joints are not required to be predicted. So the confidence of the heatmap has an additional function to predict whether the joint is in the image. On the other hand, 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 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:

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

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 3.3.

3.3 Training of the Adversarial Networks

In this section, we describe in detail how discriminators 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,

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 convergence (one may check on the validation set);
7: **return** G.

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 as negative samples. When predicted heatmaps are sufficiently close to ground-truths, considering it as a successful prediction makes sense. We also find the network to be difficult to converge by simply setting 0 or 1 as the ground truth label for a sample. Based on these observations, we design 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 vector filled with 1. 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:

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

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 that satisfy the joints constraints of human bodies.

As mentioned in Section 3.2, auxiliary confidence discriminator C is required for 2D human pose estimation. If G generates low-confidence heatmaps, C would classify the result as "fake". As G is optimized to deceive C that the fakes are 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) vector $\boldsymbol{c}_{\text{real}}$ filled with 1. 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. Therefore the fake (predicted) heatmaps are labeled with a 16×1 vector $\boldsymbol{c}_{\text{fake}}$ where the elements of $\boldsymbol{c}_{\text{fake}}$ are the corresponding confidence scores.

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

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 [59]. For our task, it is clear that we also need to supervise G in the training process with the ground truth 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.

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

Here $\alpha = 0$ if $c_{\text{fake}} = c_{\text{real}}$, $\beta = 0$ if $p_{\text{fake}} = p_{\text{real}}$. 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. When training tasks without C, we can simply set α to 0.

4 EXPERIMENTS

We evaluate the effectiveness of the proposed adversarial learning strategy on three structural tasks: 2D facial landmark detection, 2D single human pose estimation and 2D to 3D human pose transformation.

4.1 Facial Landmark Detection

Datasets. There are different strategies of annotating landmarks in the literature, such as 5 key points [53], 21 key points [66], 29 key points [67] and 68 key points [68]. We follow the 68-points annotating as the main experimental setting as the level of difficulty of estimation increases with more landmarks. The annotations are provided for LFPW [67], HELEN [69], AFW [70] and IBUG [68] datasets.

The details of these datasets are as follows: *(i)* 811 training images and 224 testing images in LFPW, *(ii)* 2000 training images and 330 testing images in HELEN, *(iii)* 337 images in AFW, *(iv)* 135 images in IBUG. These databases are used for training of our method. As the official test set of 300W competition [68] was not released at first, the testing images in LFPW and HELEN is commonly refereed as the *common test set of 300W competition [68]*, the images in IBUG is commonly refereed as the *challenging test set of 300W*. The common and challenging test sets together are referred as the *full test set of 300W*. After the later version of 300W competition, the official test set consisting of 300 indoor and 300 outdoor images was released, which was reported to have similar configuration as the IBUG dataset. In our method, we follow the standard routine to use images in LFPW, HELEN, AFW and IBUG for training and 600 official test images for testing. All annotations and bounding boxes are available at https://ibug.doc.ic.ac.uk/resources/facial-point-annotations/.

Furthermore, we conduct an ablation experiment on the AFLW dataset [66] with 21 landmarks since it contains more non-frontal faces. We follow the experimental settings in [71], [72] where landmarks of two ears are not estimated. As in [72], the dataset is split into two sets: AFLW-Full and AFLW-Frontal. The full set contains 20,000 training faces and

Table 1
Comparisons of mean error, AUC and failure rate (at a threshold of 0.08 of the normalized error) on the 300W test dataset.

Methods	Mean error (%)	AUC	Failure (%)
ESR [46]	8.47	26.09	30.50
ERT [74]	8.41	27.01	28.83
LBF [51][1]	8.57	25.27	33.67
Yan *et al.* [75]	-	34.79	12.67
Face++ [76]	-	32.23	13.00
SDM [50]	5.83	36.27	13.00
CFAN [54]	5.78	34.78	14.00
CFSS [77]	5.74	36.58	12.33
MDM [2]	4.78	45.32	6.80
DAN [3]	4.30	47.00	2.67
Baseline	**4.25**	**50.06**	**2.67**
Ours	**3.96**	**53.64**	**2.50**

[1]The implementation uses the fast version of LBF.

Table 2
Results on the AFLW facial landmark detection test set.

Methods	SDM	ERT	LBF	CFSS	SAN [78]	Baseline	Ours
AFLW-Full	4.05	4.35	4.25	3.92	1.91	1.81	**1.39**
AFLW-Front	2.94	2.75	2.74	2.68	1.85	1.67	**1.32**

4,386 testing faces. The frontal set uses the same training set but only uses 1,165 frontal faces for evaluation.

Experimental Settings. According to the estimated bounding boxes of faces, we use the center location and the diagonal distance of the bounding box to crop the face images into similar scales at the resolution of 256×256 pixels. To make the network robust to different face initialization, we follow the popular routine [2], [54] to augment samples by (0.75-1.25) scaling and $\pm25°$ in-plane rotations generated from a uniform distribution. To reduce computation consumption, the network starts with a 7×7 convolutional layer with stride 2 to downsize the resolution from 256×256 to 128×128. Then the proposed network is connected to the 128 feature maps. The networks is stacked four times in this task. For implementation, we train all our 2D pose models with the Torch7 toolbox [73].

4.1.1 Quantitative Results

We follow the same protocol of reporting errors as the 300w competition, where the average point-to-point Euclidean error normalized by the inter-ocular distance (measured as the Euclidean distance between the outer corners of the eyes) is used as the error measure.

First, we report our results in the form of CED curves in Fig. 6 which is consistent with [68]. Our method is compared against a few state-of-the-art methods, including Deep Alignment Network (DAN) [3], Mnemonic Descent Method [2], Coarse-to-Fine Shape Searching (CFSS) [77], Coarse-to-Fine Auto-encoder Networks (CFAN) [54], Local Binary Features (LBF) [51], Explicit Regression Trees (ERT) [74], Supervised Descent Method (SDM) [50], Explicit Shape Regression (ESR) [46], Deng *et al.* [79], Fan *et al.* [80], Martinez *et al.* [81], Uricar*et al.* [81], Face++ [76] and Yan *et al.* [75].

Results of the last six methods as listed are quoted from the 300W competition website. SDM is implemented by our-
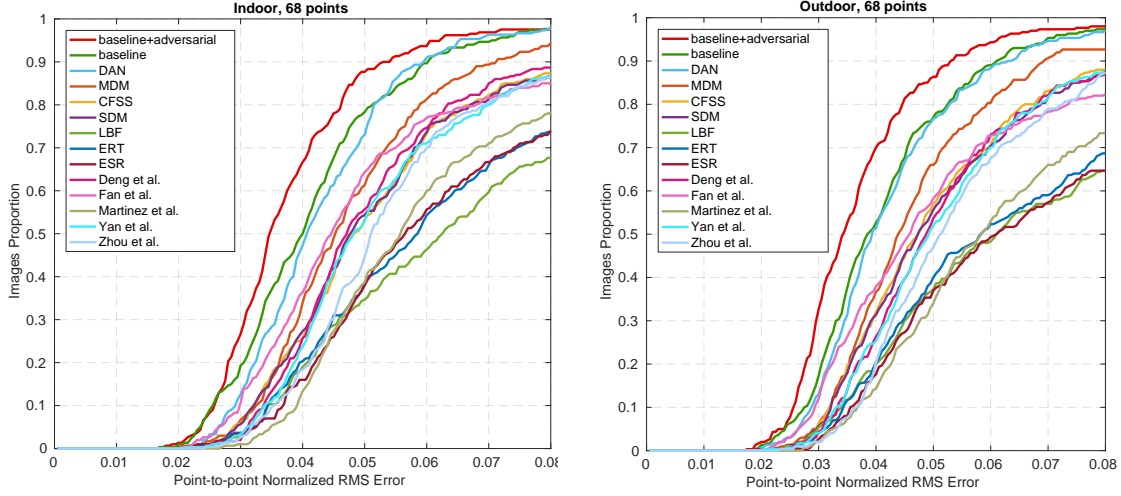
Figure 6. Quantitative results on the test set of the 300W competition (indoor and outdoor) for 68-point prediction.The point-to-point error is normalized by the inter-ocular distance.
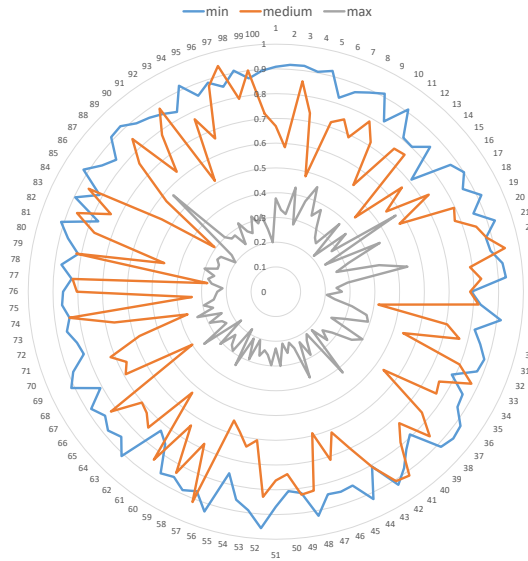

Figure 7. Results of the discriminator network *P* for the task of facial landmark estimation. The samples are sorted from the highest NRMSE error to the lowest one. The discriminator scores of the top 100 samples are marked with the gray line. The medium 100 samples are marked with the orange line. The lowest 100 samples are marked with the blue line.

self using the dense-SIFT feature provided by the author of the original paper. For other methods, publicly available implementations are used for testing. The results demonstrate that our method outperforms compared face alignments methods in every error metrics. The adversarial learning strategy clearly improves the performance compared to the baseline model.

It should be noted that, although our method avoids the coarse-to-fine approaching strategy, we perform much better in fine estimation. In particular, compared to MDM, which uses the CNN features with a recurrent process to replace the original cascaded modules, our method uses stacked modules instead and achieves better results. Compared to the insistence of cascaded strategy before, this sets a new point of view that CNN is capable of end-to-end learning such a

complex and accurate regression function for face alignment, if the network's capacity is sufficient, and more importantly, we exploit supervision appropriately for training.

We have calculated a few more metrics from the CED curve to provide insights into the performance of our method, such as mean error, area-under-the-curve (AUC) and the failure rate (at a threshold of 0.08 of the normalized error) of each method. Only the top three performing methods of the last competition are shown in the table, as in Table 1. It can be shown that, although our method improves marginally in error rate when the threshold is set at 0.08, our method greatly reduces the mean error and improves the AUC performance significantly.

For AFLW, we use the face size to normalize the mean error as the evaluation metric. The performance is reported in Table 2. Our method is compared against some of the methods in Table 1 and the Style Aggregated Network (SAN) [78]. The results of other methods on AFLW are quoted from [78]. We follow the same train/valid split and show better performance than other methods. Besides, the adversarial learning strategy again clearly improves the performance of our baseline model especially for *non-frontal faces*.

4.1.2 Qualitative Comparisons

To intuitively show the improvement of our method over previous methods, we show samples with large errors using previous methods in Fig. 8. It can be easily observed that *our method estimates more reasonable face shapes under extreme poses and occlusions.* For example, in the second column, CFSS and SDM fail to locate most of the landmarks which produce a set of disordered points. Although MDM succeeds to locate the landmarks without occlusions, it fails in the part of occluded mouth and surrounding face contour. Especially for the face contour, the landmarks are estimated without explicitly enforcing shape constraints. On the other side, our method succeeds in locating the landmarks accurately and maintains plausible face shapes.

A comparison to the baseline model is also given with faces of large occlusions in Fig. 1. We observe that the baseline model can accurately locate most landmarks under

Table 3
Comparisons of PCK@0.2 performance on the LSP dataset.

Methods	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
B.&A.'17 [84]	95.2	89.0	81.5	77.0	83.7	87.0	82.8	85.2
Lifshitz'16 [64]	96.8	89.0	82.7	79.1	90.9	86.0	82.5	86.7
Pishchulin'13 [21]	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1
Insafutdinov'16 [25]	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
Pishchulin'16 [24]	97.8	92.5	87.0	83.9	91.5	89.9	87.2	90.1
Wei'16 [8]	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
B.&T.'16 [10]	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Chou'17 [85][1]	98.2	94.9	92.2	89.5	94.2	95.1	94.1	94.0
Ours	**98.5**	**94.0**	**89.8**	**87.5**	**93.9**	**94.1**	**93.0**	**93.1**

[1]Published after the submission of our conference version.

occlusions. However, as geometric information is lacking, a few landmarks are clearly implausible. Our adversarial learning strategy can fix this problem.

To further show the usefulness of the discriminator network, we display the result scores in Fig. 7. As the generator network has been trained to successfully "deceive" the discriminator, the estimates of the final network are fairly accurate, which corresponds to a low failure rate on the 300W test set. Discrimination results for these estimations are mostly be extremely high, which would not help in terms of observing the usefulness of P.

Hence, we use a non-fully-converged intermediate generator network for evaluation. As the test set of 300W only contain 600 images, to show the results more clearly, we use another divided database: 300VW [52], [82], [83] for evaluation. We uniformly sample 4,397 images from the original video images. The intermediate generator network is used to estimate the landmark predictions. Then the predictions are sent into the final discriminator network to get the discrimination scores. In Fig. 8, we clearly observe that the low scores well correspond to the predictions with large errors, while the high scores correspond to the ones with small errors. This indicates the discrimination capability of our designed discriminator. As long as the generator successfully "deceive" this discriminator, the landmark estimations become more accurate.

4.2 2D Human Pose Estimation

Datasets. We evaluate the proposed method on three widely used benchmarks on pose estimation, *i.e.*, extended Leeds Sports Poses (LSP) [86], MPII Human Pose [87] for single-person human pose estimation and MSCOCO Keypoints dataset [88] for multi-person human pose estimation.

The LSP dataset consists of 11,000 training images and 1,000 testing images from sports activities. The MPII dataset consists of around 25,000 images with 40,000 annotated samples (about 28,000 for training, 11,000 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 3,000 samples [5], [9].

For the MSCOCO keypoint estimation dataset, 17 body joints are annotated with around 15,000 subjects. As our method doesn't involve any human detection module, we base our experiments on fixed detection results by an object

Table 4
Results on MPII Human Pose (PCKh@0.5).

Methods	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Tompson'14 [4]	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
Carreira *et al.* [90]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Tompson'15 [5]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
H.&R.'16 [91]	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Pishchulin'13 [21]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Lifshitz'16 [64]	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Gkioxari'16 [92]	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Rafi'16 [93]	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Insafutdinov'16 [25]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Wei'16 [8]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
B.&T.'16 [10]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Newell'16 [9]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Yang'17 [26]	**98.5**	96.3	91.9	88.1	**90.6**	88.0	85.0	91.5
Chou'17 [85][4]	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
Yang'17 [94][5]	98.4	96.5	91.9	88.2	91.1	88.6	85.3	91.8
Ke'18 [95][6]	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
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.
[4]The version using the same training set as our method and [5].
[4,5,6]Published after the submission of our conference version.

detector algorithm based on FPN [89] provided by [34]. Our method is trained on the MSCOCO 2017 training set and MPII multi-person dataset. The performance is tested on the MSCOCO 2017 test-dev set. All datasets provide the visibility of body parts, which are 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 [9] 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 [8], [25].

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.

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 2 and 3 days on a Tesla M40 GPU).

Next, we perform analysis on single-person human pose estimation and results for multi-person pose estimation are provided in the ablation study.
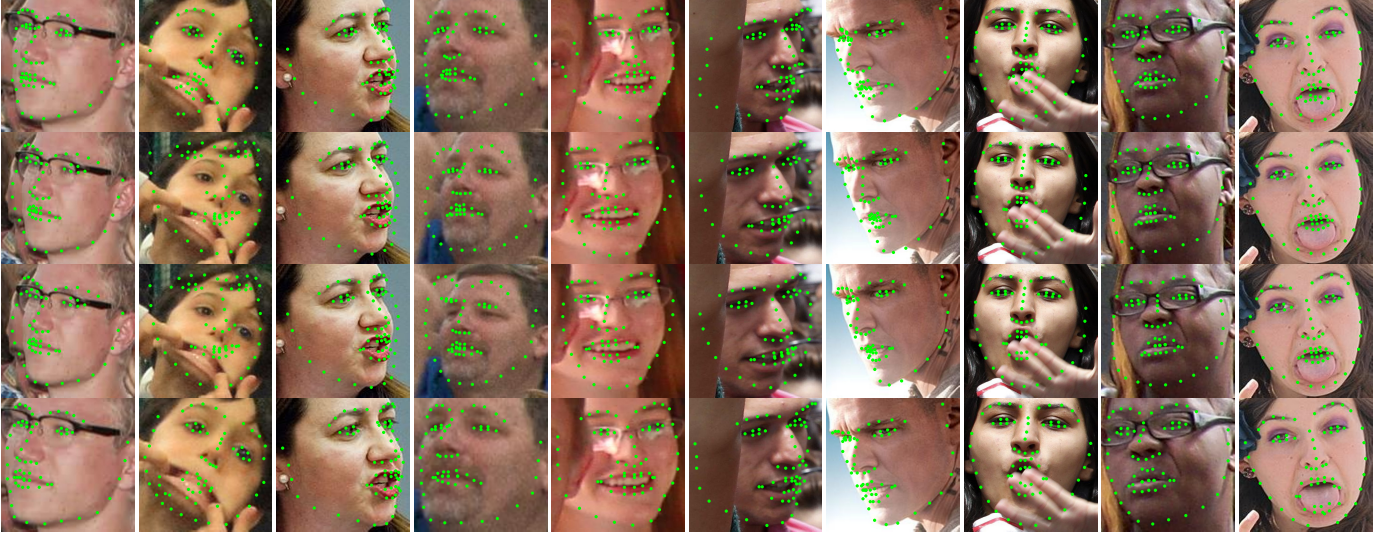
Figure 8. Samples on the 300W test set. The four rows are results of MDM [2], CFSS [77], SDM [50] and our method respectively. After estimation by each method, the coordinates are projected to the original image. Then the images are cropped to make sure that all the estimated landmarks are within the displayed image, which results in different scales of the displayed images.
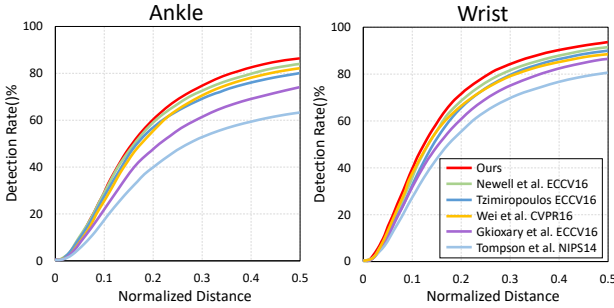

Figure 9. PCKh comparison on MPII validation set.

4.2.1   Quantitative Results

We use the Percentage Correct Keypoints (PCK@0.2) [96] 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 [87] (referred to as PCKh@0.5).

LSP Human Pose. Table 3 shows the PCK performance of our method and some existing methods at a normalized distance of 0.2. Our method achieves the second best performance, and obtains 2.4% improvement over previous methods in average. In [85] the authors also use adversarial training to improve the performance based on the hourglass network. However, it uses the auto-encoder architecture for the discriminator and uses the reconstruction loss instead of classification loss compared to our method. Nevertheless, it shows the effectiveness of adversarial training for the task of pose estimation.

MPII Human Pose. Table 4 and Fig. 9 report the PCKh performance of our method and previous methods at a normalized distance of 0.5. The baseline model here 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 [9]. 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.

Note that, recently, after the submission of this manuscript, the performance on this dataset has been further improved. Ke *et al.* [95] also emphasize on the importance of structural learning and design an structural heatmap loss, reporting a PCKh score of 92.1% on the test set. The work of [94] uses a pyramid structure to enrich the DCNN features in scale and achieves similar PCKh scores with our method.

4.2.2   Qualitative Comparisons

To gain insights on how the proposed method accomplishes 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) [9], as demonstrated in Fig. 10. For fair comparison, we also use a 2-stacked network as baseline for this experiment. We can see that our method indeed takes the structure information of the human body into consideration, leading to plausible 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 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, it may locate some body parts to the surrounding area instead of the right one. In (c), due to the lack of body configuration constraints, HG produces poses with strange twisting across body limbs. As we have implicitly embedded the body constraints into our discriminator, our network manages to predict the correct body location even under some difficult situations.

We also show some failure examples of our method on the MPII test set in Fig. 11. As shown in Fig. 11, our method
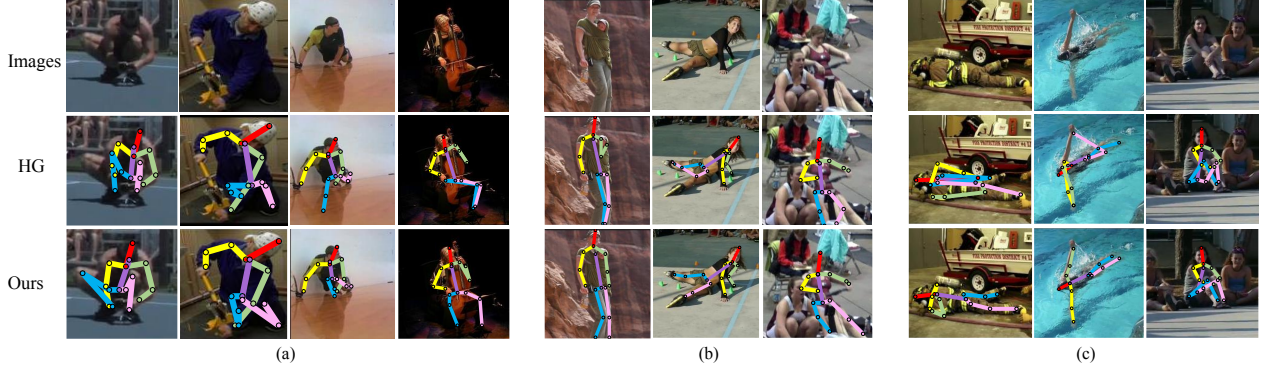
Figure 10. Prediction samples on the MPII test set. First row: original images. Second row: results by stacked hourglass network (HG) [9]. Third row: results by our method. (a)-(c) show three different failure cases of HG.

Figure 11. Failure cases caused by body parts at the edge (first and 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 [9], respectively.

Table 5
Detection rates of visible and invisible elbows and wrists.

Methods	Visible		Invisible	
	Wrist	*Elbow*	*Wrist*	*Elbow*
[9]	93.6	95.1	67.2	74.0
Ours	**94.5**	**95.9**	**70.7**	**77.6**

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 still achieves more reasonable poses compared to previous methods. Previous method may generate some poses which violate human body structure as shown in the first row of Fig. 11. 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 strange poses.

4.2.3 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 and MSCOCO keypoint dataset. Analysis about occlusion, multi-task learning and discriminators are given as follows.
Occlusion Analysis Here we present a detailed analysis of the outputs of the networks when joints in the images are

Table 6
Results on the MSCOCO keypoint detection test-dev set.

Methods	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}
CMU-Pose [32]	61.8	84.9	67.5	57.1	68.2
G-RMI [97]	68.5	87.1	75.5	65.8	73.3
Mask R-CNN [31]	63.1	87.3	68.7	57.8	71.4
Megvii [34]	72.1	91.4	80.0	68.7	77.2
RMPE [33]	61.8	83.7	69.8	58.6	67.6
RMPE++ [33]	72.3	89.2	79.1	68.0	78.6
Baseline	68.4	86.5	74.7	63.6	75.7
Ours	70.5	88.0	76.9	66.0	77.0

occluded. First, two examples with some body parts occluded are given in Fig. 12. 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 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.

For the second image, half part of the person is overlapped by the person ahead of him. Our method again manages to predict the correct pose locations with clear heatmaps. Occlusion information is also well predicted for the occluded parts. As shown with the bars in red, although the confidence scores of the occluded body parts are relatively low, they remain an overall high level. This shows that our network has learned some degree of human body priors during training. Thus it has the ability to predict plausible 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 [9] 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 5. 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 invisible wrists and elbows. This shows the advantage of our method in dealing with body parts with occlusions.*
Multi-task. We compare the four-stacked multi-task generator with the single-task model. The networks are trained

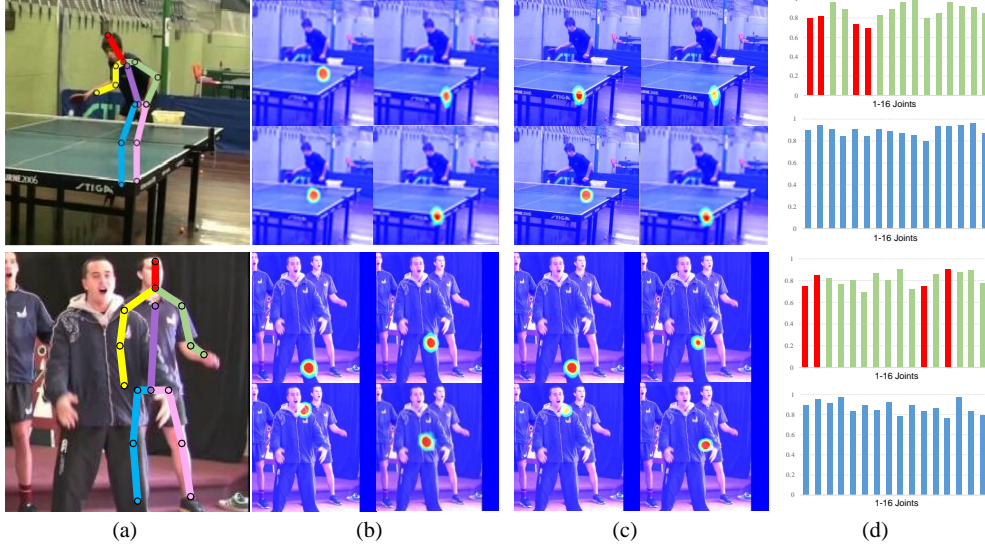|			
(a)	(b)	(c)	(d)

Figure 12. (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 bars in the output of C correspond to values of the four occluded body parts.

Table 7
Results on Human3.6M under Protocol #1 (no rigid alignment in post-processing). SA indicates that a model was trained for each action, and MA indicates that a single model was trained for all actions. For 3d baseline and our method, SH indicates that the 2D poses are estimated using the Stacked Hourglass Network, GT indicates that the ground-truth 2D poses are used. As using ground-truth 2D pose is not fair for comparison with other methods, it is only used for evaluation of PoseNet over 3D baseline.

Methods	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitingD	Smoke	Wait	WalkD	Walk	WalkT	**Mean**
LinKDE [98](SA)	132.7	183.6	132.3	164.4	162.1	205.9	150.6	171.3	151.6	243.0	162.1	170.7	177.1	96.6	127.9	162.1
Li *et al.* [99] (MA)	-	136.9	96.9	124.7	-	168.7	-	-	-	-	-	-	132.2	70.0	-	-
Tekin *et al.* [100](SA)	102.4	147.2	88.8	125.3	118.0	182.7	112.4	129.2	138.9	224.9	118.4	138.8	126.3	55.1	65.8	125.0
Zhou*et al.* [91] (MA)	87.4	109.3	87.1	103.2	116.2	143.3	106.9	99.8	124.5	199.2	107.4	118.1	114.2	79.4	97.7	113.0
Tekin *et al.* [40] (SA)	-	129.1	91.4	121.7	-	162.2	-	-	-	-	-	-	130.5	65.8	-	-
DeepViewPnt [101] (SA)	80.3	80.4	78.1	89.7	-	-	-	-	-	-	-	-	-	95.1	82.2	-
Du *et al.* [102] (SA)	85.1	112.7	104.9	122.1	139.1	135.9	105.9	166.2	117.5	226.9	120.0	117.7	137.4	99.3	106.5	126.5
Park *et al.* [103] (SA)	100.3	116.2	90.0	116.5	115.3	149.5	117.6	106.9	137.2	190.8	105.8	125.1	131.9	62.6	96.2	117.3
Zhou *et al.* [38] (MA)	91.8	102.4	96.7	98.8	113.4	125.2	90.0	93.8	132.2	159.0	107.0	94.4	126.0	79.0	99.0	107.3
Pavlakos *et al.* (MA) [41]	67.4	71.9	66.7	69.1	**72.0**	**77.0**	65.0	68.3	83.7	**96.5**	71.7	65.8	74.9	59.1	63.2	71.9
Wei *et al.*(MA) [44][1]	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	43.6	60.1	47.7	58.6
3d baseline (SH)(MA) [43]	53.3	60.8	62.9	62.7	86.4	82.4	57.8	58.7	81.9	99.8	69.1	63.9	67.1	50.9	54.8	67.5
Ours (SH)(MA)	**49.1**	**58.8**	**56.9**	**60.2**	83.0	80.1	**53.1**	**57.2**	**80.5**	96.5	**68.5**	**61.9**	**66.2**	**47.8**	**53.8**	**64.9**
3d baseline (GT)(MA) [43]	37.7	44.4	40.3	42.1	48.2	54.9	44.4	42.1	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
Ours (GT)(MA)	36.2	43.8	40.2	40.2	47.5	54.2	41.7	41.2	53.6	57.0	44.7	45.1	46.1	36.1	40.0	44.5

[1]Published after the submission this manuscript.

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 single-task model as shown in Fig. 14. This shows that the multi-task structure helps the network to understand the poses.

Discriminators. We first 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 single-task model as in

Fig. 14.

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 increases 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%. Occlusion information can clearly help understand the image and generate more accurate poses.

Second, we conducted experiments on the MPII validation set to show individual effects of P and C. This is done by simply removing P and C separately in our method.

Table 8
Results on Human3.6M under Protocol #2 (rigid alignment in post-processing). SA indicates that a model was trained for each action, and MA indicates that a single model was trained for all actions. SH indicates that the 2D poses are estimated using the Stacked Hourglass Network.

Methods	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitingD	Smoke	Wait	WalkD	Walk	WalkT	Mean
Akhter & Black [104] (SA)	199.2	177.6	161.8	197.8	176.2	186.5	195.4	167.3	160.7	173.7	177.8	181.9	176.2	198.6	192.7	181.1
Ramakrishna *et al.* [105] (MA)	137.4	149.3	141.6	154.3	157.7	158.9	141.8	158.1	168.6	175.6	160.4	161.7	150.0	174.8	150.2	157.3
Zhou *et al.* [106] (SA)	99.7	95.8	87.9	116.8	108.3	107.3	93.5	95.3	109.1	137.5	106.0	102.2	106.5	110.4	115.2	106.7
Bogo [39] (MA)	62.0	60.2	67.8	76.5	92.1	77.0	73.0	75.3	100.3	137.3	83.4	77.3	86.8	79.7	87.7	82.3
Moreno-Noguer [42] (SA)	66.1	61.7	84.5	73.7	65.2	67.2	60.9	67.3	103.5	74.6	92.6	69.6	71.5	78.0	73.2	74.0
Pavlakoset al. [41] (SA)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	51.9
3d baseline (SH)(MA) [43]	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Wei *et al.* (MA) [44][1]	26.9	30.9	36.3	39.9	43.9	47.4	28.8	29.4	36.9	58.4	41.5	30.5	29.5	42.5	32.2	37.7
Ours (SH)(MA)	**38.5**	**42.7**	**43.9**	**46.1**	**49.1**	**53.2**	**41.0**	**39.8**	**53.9**	**63.8**	**48.1**	**43.9**	**49.3**	**37.6**	**41.0**	**46.1**

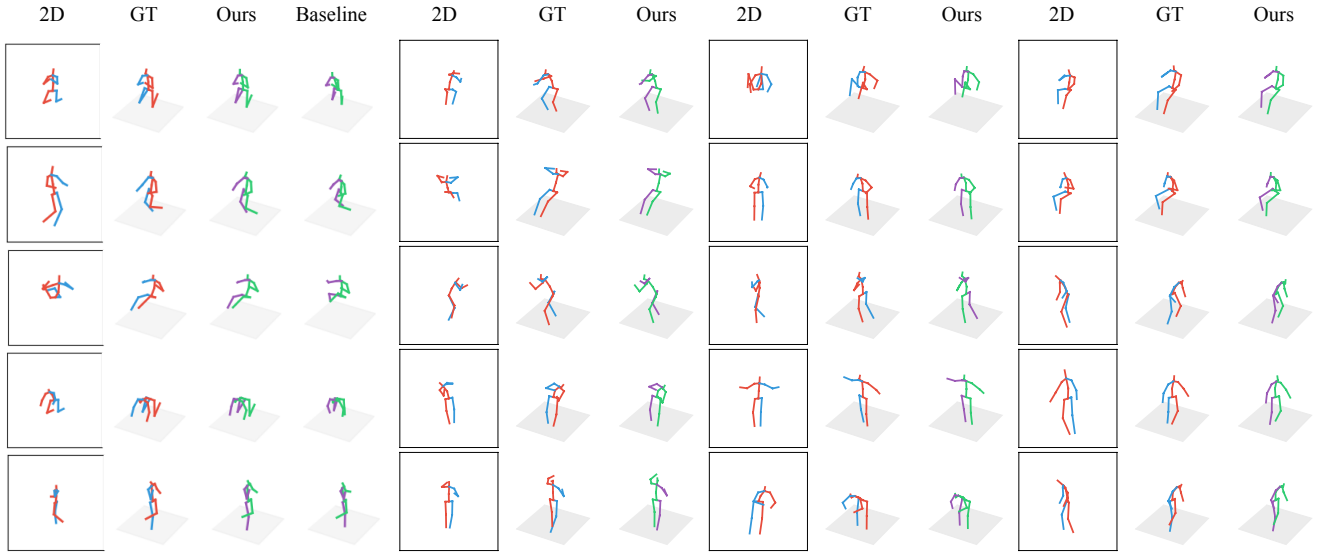[1]Published after the submission of this manuscript.

Figure 13. 3D pose examples on the Human 3.6M validation set. The left four columns show comparisons with baseline model. The proposed adversarial learning method refines the implausible poses generated by the baseline model and produces results more similar to ground-truth poses (GT). Other columns show 3D poses generated by our method in different scenes.

With single P, the performance of our method evaluated by PCKh@0.5 is 91.9% compared to 91.1% by the baseline. With single C, the performance is 91.4%. It is clear that P contributes more to our final improvement. P incorporates information of whether the pose configuration is plausible. C uses the same ℓ_2 loss as baseline while using an adversarial learning strategy.

Multi-person In the problem of multi-person pose estimation, occlusions and overlappings are more serious than single-person pose estimation. Although our method focuses on single-person human pose estimation, we conduct ablation experiment on the MSCOCO Keypoints dataset [88] to validate the effectiveness of our method under these circumstances. The results are displayed in Table 6. The performance of other methods is quoted from [33]. The adversarial learning strategy improves the performance of the baseline model.
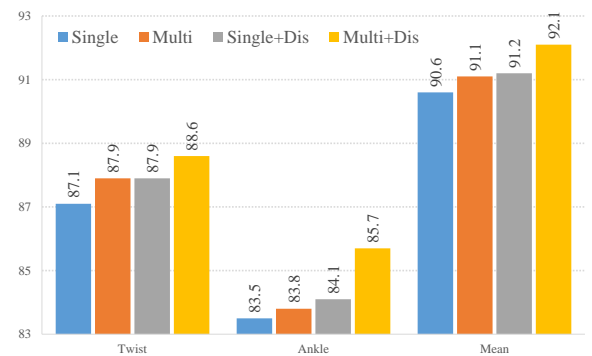

Figure 14. Ablation study: PCKh scores at the threshold of 0.5.

4.3   2D to 3D Pose Transformation

Datasets and Experimental Settings We focus our numerical evaluation on a public datasets for 3d human pose estimation: Human 3.6M [98]. Human 3.6M is currently the largest

publicly available datasets for human 3d pose estimation. The dataset consists of 3.6 million images featuring 7 professional actors performing 15 everyday activities such as walking, eating, sitting, making a phone call and engaging in a discussion. 2d joint locations and 3d ground truth positions are available, as well as projection (camera) parameters and body proportions for all the actors. We follow the standard protocol, using subjects 1, 5, 6, 7, and 8 for training, and subjects 9 and 11 for validation. For fair comparison to previous methods, we build our method based on a recently published baseline [43] and strictly follow their experimental settings.

In detail, the 2D to 3D transformation net is the same as [43]. 2D poses are estimated by the same hourglass networks as [43]used. We only add our discriminators and adversarial training to provide structural information to the original method. The average error in millimeters between the ground truth and our prediction across all joints and cameras are reported, after alignment of the root (central hip) joint. In some of the baselines, the prediction has been further aligned with the ground truth via a rigid transformation (*e.g.*, [39], [42]). We refer the experiment without further alignment as Protocol # 1 while the one with alignment as Protocol #2. On the other hand, some recent methods have trained one model for all the actions, as opposed to building action-specific models instead of independent training and testing in each action. We also show their results under these two circumstances.

Table 7 reports the results without further alignment and Table 8 reports the results with further alignment. By simply adding a structural PoseNet structure on [43], the performance is improved. The work of [44] followed the same adversarial learning routine with our method and proposed more complex and discriminators to tell whether its prediction is plausible. It demonstrates that *the performance of adversarial learning framework can be further improved by designing better discriminators.* It should be pointed out all the gain in performance comes with no additional computation cost during test.

We also show examples on Human 3.6M of both the baseline model and the proposed method to show the effectiveness of adversarial learning in Fig. 13. Some implausible 3D poses generated by the baseline model is well refined by our method.

5 CONCLUSIONS

In this paper, we have proposed a novel conditional adversarial network for pose estimation, which trains a pose generator with discriminator networks. The discriminators function as an expert who distinguishes plausible poses from unreasonable ones. By training the pose generator to deceive the expert that the generated pose is real, our network is more robust to occlusions, overlapping and twisting of pose components. In contrast to previous work using DCNNs for pose estimation, our network is able to alleviate the risk of localizing human body parts onto the matched features without consideration of human body priors.

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

The method developed here can be immediately applied to other shape estimation problems using DCNNs with minimal modification. The inputs of the discriminators can also be further improved to boost the discrimination ability. More significantly, we believe that the use of conditional GANs as a tool to predict structured output or enforce output dependency can be further developed to much more general structured output learning.

Yu Chen received the BS degree in mathematics and applied mathematics from Nanjing University of Science and Technology. He is currently working towards the PhD degree in the same university. He is also a researcher of Motovis Research Australia. His current research interests are deep learning, autonomous driving and pose estimation in particular.

Chunhua Shen is a Professor at School of Computer Science, University of Adelaide. Before he moved to Adelaide, he was with the computer vision program at NICTA (National ICT Australia), Canberra Research Laboratory for about six years. He studied at Nanjing University, at Australian National University, and received his PhD degree from the University of Adelaide. From 2012 to 2016, he held an Australian Research Council Future Fellowship.

Hao Chen received the master's degree from Zhejiang University, China. He is working towards the PhD degree at School of Computer Science, The University of Adelaide. His current research interests in deep learning and its applications in computer vision and text analysis.

Xiu-Shen Wei received the BS degree in computer science, and his PhD degree in computer science from Nanjing University. He is currently the Research Lead of Megvii (Face++) Nanjing Research. He achieved the first place in the Apparent Personality Analysis competition (in association with ECCV 2016) and the first runner-up in the Cultural Event Recognition competition (in association with ICCV 2015) as the team director. He also received the Presidential Special Scholarship (the highest honor for Ph.D. students) in Nanjing University. His research interests are computer vision and machine learning.

Lingqiao Liu received the BS and MS degrees in communication engineering from the University of Electronic Science and Technology of China, Chengdu, in 2006 and 2009, respectively, and the PhD degree from the Australian National University, Canberra, in 2014. He is now a Lecturer at the University of Adelaide. In 2016, he was awarded the Discovery Early Career Researcher Award by the Australian Research Council. His research interests include various topics in computer vision and machine learning.

Jian Yang received the PhD degree from Nanjing University of Science and Technology (NUST), on the subject of pattern recognition and intelligence systems in 2002. In 2003, he was a postdoctoral researcher at the University of Zaragoza. From 2004 to 2006, he was a Postdoctoral Fellow at Biometrics Centre of Hong Kong Polytechnic University. From 2006 to 2007, he was a Postdoctoral Fellow at Department of Computer Science of New Jersey Institute of Technology. Now, he is a Chang-Jiang professor in the School of Computer Science and Technology of NUST. He is the author of more than 100 scientific papers in pattern recognition and computer vision. His journal papers have been cited more than 4000 times in the ISI Web of Science, and 9000 times in the Web of Scholar Google. His research interests include pattern recognition, computer vision and machine learning. Currently, he is/was an associate editor of Pattern Recognition Letters, IEEE Trans. Neural Networks and Learning Systems, and Neurocomputing. He is a Fellow of IAPR.

REFERENCES

[1] A. Jourabloo, M. Ye, X. Liu, and L. Ren, "Pose-invariant face alignment with a single cnn," in *Proc. IEEE Int. Conf. Comp. Vis.* IEEE, 2017, pp. 3219–3228.

[2] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou, "Mnemonic descent method: A recurrent process applied for end-to-end face alignment," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016, pp. 4177–4187.

[3] M. Kowalski, J. Naruniec, and T. Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," *Proc. IEEE Conf. Comp. Vis. Patt. Recogn. Workshop*, 2017.

[4] 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.*, 2014, pp. 1799–1807.

[5] 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.*, 2015, pp. 648–656.

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

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

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

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

[10] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *Proc. Eur. Conf. Comp. Vis.*, 2016, pp. 717–732.

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

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

[13] 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.*, 2016, pp. 2226–2234.

[14] 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.*, 2014, pp. 2672–2680.

[15] 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.*, 2015, pp. 1486–1494.

[16] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *Proc. IEEE Int. Conf. Comp. Vis.*, vol. 1, no. 6, 2017, p. 8.

[17] 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*, vol. 99, no. 2, pp. 190–214, 2012.

[18] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman, "Upper body detection and tracking in extended signing sequences." *Int. J. Comput. Vision*, vol. 95, no. 2, pp. 180–197, 2011.

[19] B. Sapp, A. Toshev, and B. Taskar, "Cascaded models for articulated pose estimation," in *Proc. Eur. Conf. Comp. Vis.*, 2010, pp. 406–420.

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

[21] 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.*, 2013, pp. 3487–3494.

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

[23] 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.*, 2016, pp. 3073–3082.

[24] 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.*, 2016, pp. 4929–4937.

[25] 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.*, 2016, pp. 34–50.

[26] 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.

[27] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik, "Using k-poselets for detecting people and localizing their keypoints," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014, pp. 3582–3589.

[28] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele, "Articulated people detection and pose estimation: Reshaping the future," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* IEEE, 2012, pp. 3178–3185.

[29] M. Sun and S. Savarese, "Articulated part-based model for joint object detection and pose estimation," in *Proc. IEEE Int. Conf. Comp. Vis.* IEEE, 2011, pp. 723–730.

[30] S. Huang, M. Gong, and D. Tao, "A coarse-fine network for keypoint localization," in *Proc. IEEE Int. Conf. Comp. Vis.*, vol. 2, 2017.

[31] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proc. IEEE Int. Conf. Comp. Vis.* IEEE, 2017, pp. 2980–2988.

[32] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, vol. 1, no. 2, 2017, p. 7.

[33] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[34] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018.

[35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. Advances in Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[36] H.-J. Lee and Z. Chen, "Determination of 3d human body postures from a single view," *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 2, pp. 148–168, 1985.

[37] J. A. Scott and C. W. Binns, "Factors associated with the initiation and duration of breastfeeding: a review of the literature." *Breastfeeding review: professional publication of the Nursing Mothers' Association of Australia*, vol. 7, no. 1, pp. 5–16, 1999.

[38] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei, "Deep kinematic pose regression," in *European Conf. Computer Vision–ECCV 2016 Workshops.* Springer, 2016, pp. 186–201.

[39] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *Proc. Eur. Conf. Comp. Vis.* Springer, 2016, pp. 561–578.

[40] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua, "Structured prediction of 3d human pose with deep neural networks," *Proc. British Machine Vis. Conf.*, 2016.

[41] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3d human pose," *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.

[42] F. Moreno-Noguer, "3d human pose estimation from a single image via distance matrix regression," *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.

[43] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[44] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang, "3d human pose estimation in the wild by adversarial learning," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018.

[45] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2010, pp. 1078–1085.

[46] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *Int. J. Comput. Vision*, vol. 107, no. 2, pp. 177–190, 2014.

[47] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool, "Real-time facial feature detection using conditional regression forests," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* IEEE, 2012, pp. 2578–2585.

[48] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* IEEE, 2010, pp. 2729–2736.

[49] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2013, pp. 1513–1520.

[50] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2013, pp. 532–539.

[51] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014, pp. 1685–1692.

[52] G. Tzimiropoulos, "Project-out cascaded regression with an application to face alignment," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015, pp. 3659–3667.

[53] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2013, pp. 3476–3483.

[54] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment," in *Proc. Eur. Conf. Comp. Vis.* Springer, 2014, pp. 1–16.

[55] Y. Chen, W. Luo, and J. Yang, "Facial landmark detection via pose-induced auto-encoder networks," in *Image Processing (ICIP), 2015 IEEE International Conference on.* IEEE, 2015, pp. 2115–2119.

[56] Y. Chen, J. Qian, J. Yang, and Z. Jin, "Face alignment with cascaded bidirectional lstm neural networks," in *Pattern Recognition (ICPR), 2016 23rd International Conference on.* IEEE, 2016, pp. 313–318.

[57] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," *arXiv preprint arXiv*, vol. 1411.1784, pp. 1–7, 2014.

[58] 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.*, 2016, pp. 1–10.

[59] 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.*, 2016, pp. 2536–2544.

[60] X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," in *Proc. Eur. Conf. Comp. Vis.*, 2016, pp. 318–335.

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

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

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

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

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

[66] P. M. R. Martin Koestinger, Paul Wohlhart and H. Bischof, "Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization," in *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[67] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2930–2940, 2013.

[68] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 397–403.

[69] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proc. Eur. Conf. Comp. Vis.* Springer, 2012, pp. 679–692.

[70] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* IEEE, 2012, pp. 2879–2886.

[71] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou, "A deep regression architecture with two-stage re-initialization for high performance facial landmark detection," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.

[72] S. Zhu, C. Li, C.-C. Loy, and X. Tang, "Unconstrained face alignment via cascaded compositional learning," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016, pp. 3409–3417.

[73] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, NIPS workshop*, no. EPFL-CONF-192376, 2011.

[74] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014, pp. 1867–1874.

[75] J. Yan, Z. Lei, D. Yi, and S. Li, "Learn to combine multiple hypotheses for accurate face alignment," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 392–396.

[76] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 386–391.

[77] S. Zhu, C. Li, C. Change Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015, pp. 4998–5006.

[78] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style aggregated network for facial landmark detection," *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018.

[79] J. Deng, Q. Liu, J. Yang, and D. Tao, "M 3 csr: multi-view, multi-scale and multi-component cascade shape regression," *Image and Vision Computing*, vol. 47, pp. 19–26, 2016.

[80] H. Fan and E. Zhou, "Approaching human level facial landmark localization by deep learning," *Image and Vision Computing*, vol. 47, pp. 27–35, 2016.

[81] B. Martinez and M. F. Valstar, "L 2, 1-based regression and prediction accumulation across views for robust facial landmark detection," *Image and Vision Computing*, vol. 47, pp. 36–44, 2016.

[82] G. G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape, "Offline deformable face tracking in arbitrary videos," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 1–9.

[83] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 50–58.

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

[85] C.-J. Chou, J.-T. Chien, and H.-T. Chen, "Self adversarial training for human pose estimation," *arXiv: Comp. Res. Repository*, vol. 1707.02439, 2017.

[86] 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.

[87] 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.*, 2014, pp. 3686–3693.

[88] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comp. Vis.* Springer, 2014, pp. 740–755.

[89] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, vol. 1, no. 2, 2017, p. 4.

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

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

[92] G. Gkioxari, A. Toshev, and N. Jaitly, "Chained predictions using convolutional neural networks," in *Proc. Eur. Conf. Comp. Vis.*, 2016, pp. 728–743.

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

[94] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *Proc. IEEE Int. Conf. Comp. Vis.*, vol. 2, no. 7, 2017.

[95] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *Proc. Eur. Conf. Comp. Vis.*, 2018.

[96] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, 2013.

[97] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multiperson pose estimation in the wild," *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, vol. 8, 2017.

[98] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, 2014.

[99] S. Li, W. Zhang, and A. B. Chan, "Maximum-margin structured learning with deep networks for 3d human pose estimation," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2015, pp. 2848–2856.

[100] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua, "Direct prediction of 3d body poses from motion compensated sequences," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016, pp. 991–1000.

[101] M. F. Ghezelghieh, R. Kasturi, and S. Sarkar, "Learning camera viewpoint using cnn to improve 3d body pose estimation," in *3D Vision (3DV), 2016 Fourth International Conference on.* IEEE, 2016, pp. 685–693.

[102] Y. Du, Y. Wong, Y. Liu, F. Han, Y. Gui, Z. Wang, M. Kankanhalli, and W. Geng, "Marker-less 3d human motion capture with monocular image sequence and height-maps," in *Proc. Eur. Conf. Comp. Vis.* Springer, 2016, pp. 20–36.

[103] S. Park, J. Hwang, and N. Kwak, "3d human pose estimation using convolutional neural networks with 2d pose information," in *Computer Vision–ECCV 2016 Workshops.* Springer, 2016, pp. 156–169.

[104] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3d human pose reconstruction," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015, pp. 1446–1455.

[105] V. Ramakrishna, T. Kanade, and Y. Sheikh, "Reconstructing 3d human pose from 2d image landmarks," *Proc. Eur. Conf. Comp. Vis.*, pp. 573–586, 2012.

[106] X. Zhou, M. Zhu, S. Leonardos, and K. Daniilidis, "Sparse representation for 3d shape estimation: A convex relaxation approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1648–1661, 2017.