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

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

Abstract—Landmark/pose estimation in single monocular images have received much effort in computer vision due to its important applications. It remains a challenging task when input images 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 single human pose estimation, 2D facial landmark estimation and 3D single human pose estimation. The proposed approach significantly outperforms the state-of-the-art methods and almost always generates plausible pose predictions, demonstrating the usefulness of implicit learning of structures using GANs.

Index Terms—Landmark/Pose Estimation, Structure-aware Fully Convolutional Networks, Adversarial Generative Networks, Multi-task Learning.

1 Introduction

Landmark localization, a.k.a, 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. Understanding of a person's limb articulation location is very helpful for high-level vision tasks like human tracking, action recognition, and also serves as a fundamental tool in applications such as human-computer interaction. 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 achieve state-of-the-art on both tasks.

Recently, significant improvements have been made 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. The regression models have shown great capability of learning powerful feature representations. However, for pose components with heavy occlusions and background clutters that appear similar to body parts, DCNNs may encounter difficulty in regressing accurate poses.

Human vision is capable of learning the variety of shape structures from abundant observations. Even under extreme occlusions, one can effortlessly infer the potential poses and exclude the implausible ones. It is, however, very challenging to incorporate these 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 powerful features.

As a consequence, an unreasonable pose may be produced by conventional DCNNs. As shown in Fig. 1, taking human pose estimation as examples, in challenging test cases with heavy occlusions, 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 labelled 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" that 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

- *Y. Chen and J. Yang are with Nanjing University of Science and Technology, China; C. Shen and L. Liu are with The University of Adelaide, Australia; X.-S. Wei is with Nanjing University, China.*
- *This work was done when Y. Chen and X.-S. Wei were visiting The University of Adelaide.*

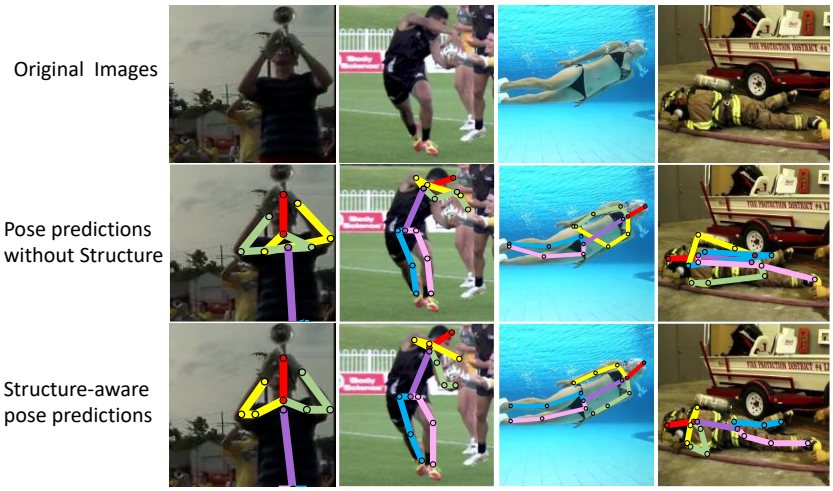


FIGURE 1: ***Motivation***. *We show the importance of strongly enforcing priors about the human body structure during training of DCNNs for pose estimation. Learning without using such priors generates inaccurate results.*

human body structure.

Inspired by the recent success in Generative Adversarial Networks (GAN) [11], [12], [13], [14], [15], we propose to design the "discriminator" as the discriminator network in GAN while the regression network functions as the generative network. Training the generator in the adversarial manner against the discriminator precisely meets our intention.

To accomplish the above goals, the discriminator should be fed with sufficient information to perform classification, while the generator should have the ability in modeling the complicated features in pose estimation. Thus, a baseline stacked bottom-up, top-down network G is designed to capture features as different spatial levels. To better capture pose structure in complex pose tasks (*e.g.*, human pose estimation), a multi-task learning network is designed, which simultaneously regresses the pose heatmaps and the occlusion heatmaps. Built upon the pose (and occlusion) heatmaps, the pose discriminator (P) is used to employed whether the pose configuration is plausible or not.

In addition, in the area of human pose estimation, our preliminary results show that correct locations often correspond to heatmaps with high confidence. Therefore, we design another discriminator C to make a decision on the confidence of the predicted pose heatmaps. The generator is asked to "fool" both the pose and confidence discriminators by training G and $\{P, C\}$ in the generative adversarial manner. Thus, the human body structure is implied in the P net by guiding G to 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.

What is more, the function of the discriminator is not limited to 2D pose estimation. For tasks concerning structured outputs (*e.g.*, 2D to 3D human pose transformation), we can also use an adversarial discriminator to learn the structure distributions to generate 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.
- We design a novel network framework for human pose estimation which takes the geometric constraints of human joints connectivity into consideration. By incorporating the priors of the human body, prediction mistakes caused by occlusions and cluttered backgrounds are considerably reduced. Even when the network fails, the outputs of the network appear more like "human" predictions instead of "machine" predictions.
- We evaluate our method on public 2D human pose estimation datasets, face landmark estimation datasets and 3D human pose estimation datasets. Our approach significantly outperforms state-of-the-art methods, and is able to consistently produce more plausible keypoint predictions compared to previous 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 cascade procedure is avoided.

1.1 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 mainly lies in detection of all people and overlapping between people. While in single-person human pose estimation tasks 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 in this work. In terms of vision tasks, our method is mostly related to

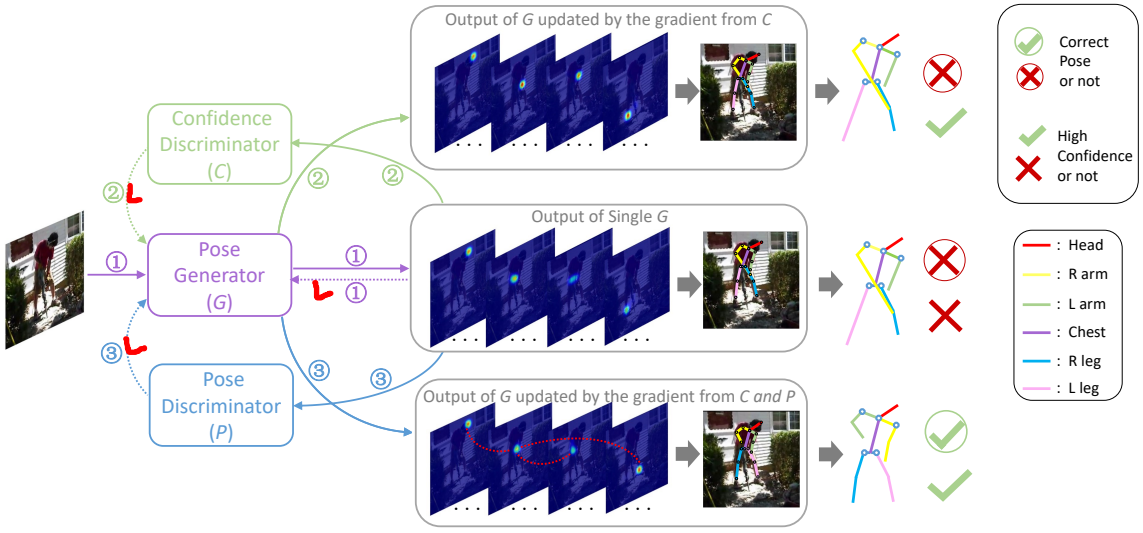


FIGURE 2: **Overview of the proposed Structure-Aware Convolutional Network for human pose estimation**. *The sub-network in purple is the stacked multi-task network (G) for pose generation. The networks in blue (P) and green (C) are used to discriminate whether the generated pose is "real" (plausible as a body shape) and whether the generator has strong confidence in locating the joints, respectively. The loss of G has three parts: mean squared error of heatmaps (dashed line in purple), adversarial loss from P (dashed line in blue) and adversarial loss from C (dashed line in green). Standalone training of G produces results shown in the middle. G and C produce results on the top while G, C and P produce results at the bottom.*

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 models [17], [18], [19], [20], [21], [22]. With the introduction of "DeepPose" by Toshev *et al*. [6], deep network based methods become increasingly popular in this area. This work is related to the methods generating pose heatmaps from images [4], [5], [7], [8], [9], [23], [24], [25]. For example, Tompson *et al*. [4] generated heatmaps by running an image through multiple resolution filter banks in parallel to simultaneously extract features at a variety of scales. Tompson *et al*. [5] used multiple branches of convolutional networks to fuse the features from an image pyramid, and used Markov Random Field (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] proposed a state-of-the-art architecture for stacked iterative inference with residual blocks. Based on the hourglass structure, Chu *et al*. [26] incorporated convolutional neural networks with a multi-context attention mechanism into an end-to-end framework for human pose estimation. The structure of our *G* net is also a fully convolutional network with "conv-deconv" architecture. However, our network is designed in a multi-task manner for improved performance.

3D Human Pose Estimation. Based on the 2D human pose prediction, 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*. [27]. As the literature of this problem is vast with approaches in a variety of settings 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 from images [28], [29]. These approaches avoid estimating 3D joint positions directly. Recently, some systems have explored the possibility of directly inferring 3D poses from images with end-to-end deep architectures [30]. Pavlakos *et al*. [31] 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 [32] learned to predict a pairwise distance matrix (DM) from 2D-to-3D space. A major motivation behind this DM regression approach, as well as the volumetric approach of Pavlakos *et al*., is that predicting 3D keypoints from 2D detections is inherently difficult. However, Martinez *et al*. [33] contradicted this idea and showed that a well-designed and simple neural network can perform competitively for the task of 2D-to-3D keypoint regression. As this network is of simple structure and achieves high performance, our method is built upon it. We also use a discriminator to determine whether the predicted 3D pose is geometrically plausible in terms of the 2D pose. We demonstrate that the idea of enforcing the adversarial PoseNet on the baseline model also works well for this 2D-to-3D problem.

2D Face Landmark Estimation. Traditional regression based methods often follow a cascade manner to update the landmark localization results in a coarse-to-fine fashion. This strategy has been proved to be very effective for face alignment. Early methods mainly use random forest regression as the regressors due to computational efficiency [34], [35], [36], [37]. Burgos-Artizzu *et al*. proposed the Robust Cascaded Pose Regression (RCPR) which increases robustness to outliers by detecting occlusions explicitly [38]. Different from the previous learning process, Supervised Descent Method (SDM) [39] 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 the handcrafted features, Local Binary Features (LBF) [40] are learned for linear regression by using the regression forest. Project-Out Cascaded Regression [41] was proposed by learning and employing a sequence of averaged Jacobian's and descent directions in a subspace orthogonal to the facial appearance variation.

Recently, deep neural networks were also introduced for face alignment [42], [43], [44], [45]. These methods use deep networks to replace the traditional regressors but still follow the cascade framework. It is worth pointing out that the Mnemonic Descent Method (MDM) [2] made face alignment end-to-end by training a convolutional recurrent neural network architecture. The original cascade steps are connected by recurrent connections and the handcrafted features are replaced by learnable convolutional features. We make 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 obtained.

Generative Adversarial Networks. Generative Adversarial Networks have been widely studied in previous work for discrete labels [46], text [47] and also images. The conditional GAN models have tackled inpainting [48], image prediction from a normal map [49], future frame prediction [50], future state prediction [51], product photo generation [52], and style transfer [53].

Human pose estimation can been 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.

2 The Proposed Adversarial PoseNet

For convenience of exposition, the network structure discussed here in this section is mainly for 2D human pose estimation. By simply removing the multi-task part and the confidence discriminator, it becomes the network for facial landmark localization that we use in this work. The training procedure is the same as for 2D human pose by setting the weight for confidence discriminator to 0.

For 2D to 3D transformation, the generator is the same as the network in [33]. As it only transforms 2D coordinates to 3D coordinates instead of from the high-dimension images to heatmaps, the discriminator uses a similar and simple structure as the generator. The training procedure is the same with facial landmark detection.

As shown in Fig. 2, our adversarial PoseNet model consists of three parts, *i.e.*, the pose generator network G, the pose discriminator network P and the confidence discriminator C. The generative network is a bottom-up and top-down network, where the inputs are the RGB images and the outputs are heatmaps for each input image (*e.g.*, 32 heatmaps for estimating 16 human joints, or 68 heatmaps for face alignment in our experiments). In the case of human pose estimation, half of the returned heatmaps are pose estimations for 16 pose key points, and the other half are for the corresponding occlusion predictions. The values in each heatmap are confidence scores normalized in the range of $[0, 1]$ where a Gaussian blur is applied around the ground truth position.

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

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

In practical training, the three parts of loss are added together to optimize G at the same time.

2.1 Multi-Task Generative Network

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

The goal of the multi-task generative network is to learn a function $\mathcal{G}$ which attempts to project an image $\boldsymbol{x}$ to both the corresponding pose heatmaps $\boldsymbol{y}$ and occlusion heatmaps $\boldsymbol{z}$, *i.e.*, $\mathcal{G}(\boldsymbol{x}) = \{\hat{\boldsymbol{y}}, \hat{\boldsymbol{z}}\}$ where $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{z}}$ are the predicted heatmaps. In addition, as reported in [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 this goal, an "encoder-decoder" architecture is used.

Besides, for the problem of human pose estimation, local evidence is essential for identifying features for human joints. Meanwhile, the final pose estimation requires a coherent understanding of the full body image. To capture this information at each scale, skip connections between mirrored layers in the encoder and decoder are added. Inspired by [9], our network is also stacked to provide the network with a mechanism for re-evaluation of initial estimates and features across the entire image. In each module of the G net, a residual block [54] is used for the convolution operator. Given the original image $\boldsymbol{x}$, a basic block of the stacked multi-task generator network can be expressed as follows:

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

where $\boldsymbol{Y}_n$ and $\boldsymbol{Z}_n$ are the output activation tensors of the n-th stacked generative network for pose estimations and occlusion predictions, respectively. $\boldsymbol{X}$ is the image feature

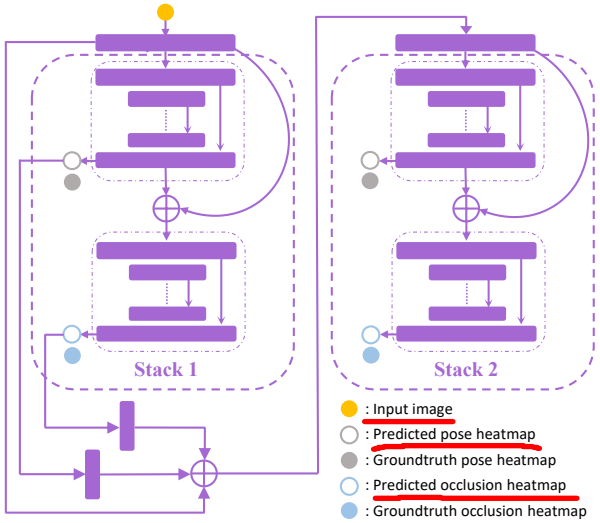


FIGURE 3: *Architecture of the multi-task generator network G. Each rectangular block indicates a convolutional block. ⊕ indicates addition of input features. The stacking of the first and the second networks is shown and more networks can be stacked with the same structure.*

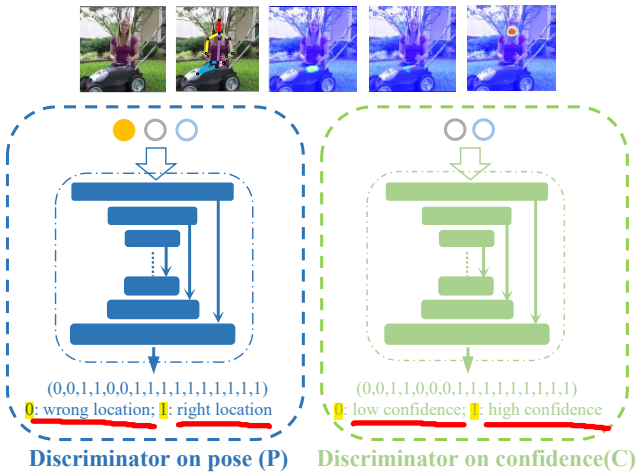


FIGURE 4: *Architectures of the discriminator network P and C. On the top we show the image for pose estimation, the image with estimated joints and heatmaps of right ankle, pelvis and neck (1st, 7th and 9th of all pose heatmaps respectively). The expected output for this sample is given in the bottom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

tensor, obtained after pre-processing on the original image through two residual blocks. Suppose that there are N times stacking of the basic block, then the multi-task generative network can be formulated as:

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

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

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

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

where Θ denotes the parameter set.

2.2 Pose Discriminator

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

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

Additionally, even when the generative network fails to predict the correct pose locations for a particular image, the predicted pose may still be a plausible one for another human body shape. Thus, simply using the pose and occlusion features may still face difficulty in training an accurate P. *Such inference should be made by taking the original image into consideration at the same time.* Occlusion information can also be useful in inferring the pose rationality. So we use the input RGB image with pose and occlusion heatmaps generated by the G net as the input to P for predicting whether a pose is reasonable or not for a particular image. The network structure of P is shown in Fig. 4. To achieve this goal, GAN is employed in the conditional manner for P in our framework. As GANs learn a generative model of data, conditional GANs (cGANs) learn a conditional generative model [14]. The objective of a conditional adversarial P network is expressed as follows:

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

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

2.3 Confidence Discriminator

By observing the differences between ground-truth heatmaps and predicted heatmaps generated 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 "guess" the body parts. This is mainly because we already acquire the geometric priors of human body joints. Motivated by this, we design a second auxiliary discriminator, which is termed Confidence

Algorithm 1 The training process of our method.

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

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 2.4.

2.4 Training of the Adversarial Networks

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

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

However, in previous works on image translation using GANs, the discriminative network is learned with all fake samples being labeled 0. When predicted heatmaps are close enough to ground-truth, 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 ground truth 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 unit vector. For the fake samples, if a predicted body part is far from the ground truth location, the pose is clearly implausible for the body configuration in this image. Therefore, when training P, the ground truth $\boldsymbol{p}_{\text{fake}}$ is:

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

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

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

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

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

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

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

Previous approaches to conditional GANs have found it beneficial to mix the GAN objective with a traditional loss, such as ℓ_2 distance [48]. For our task, it is clear that we also need to supervise G in the training process using the ground-truth human poses. Thus, the discriminator still plays the original role, but the generator will not only fool the discriminator but also approximate the ground-truth output in an ℓ_2 sense as in Eq. (3). Therefore, the final objective function is presented as follows.

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

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

3 Experiments

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

3.1 Facial Landmark Detection

Datasets. There are different strategies of annotating landmarks in the literature, such as 5 key points [42], 29 key points [55] and 68 key points [56], [57]. We follow the 68-points annotating as it provides the essential shape information in most circumstances. The annotations are provided for LFPW [55], HELEN [58], AFW [59] and IBUG [56] 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 [56] was not released at first, the testing images in

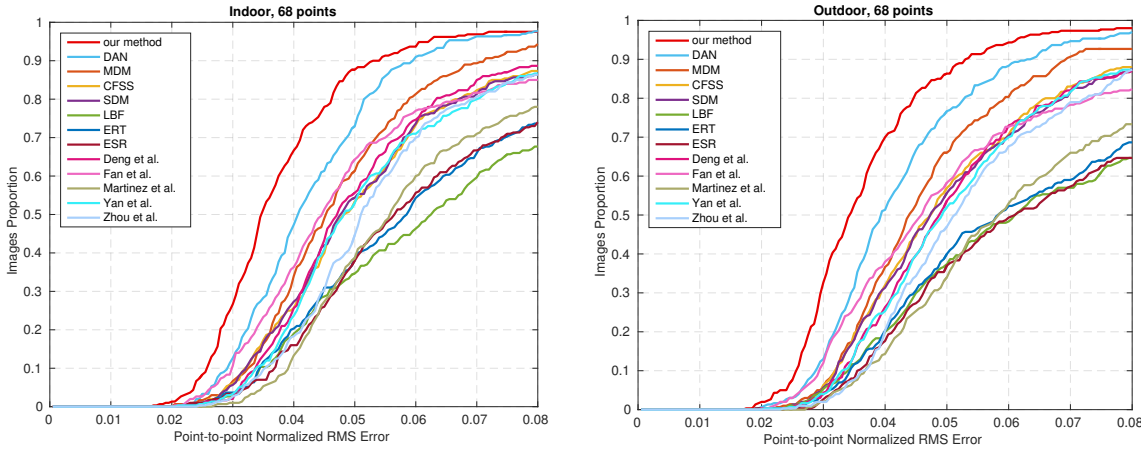


FIGURE 5: *Quantitative results on the test set of the 300W competition (indoor and outdoor) for 68-point prediction.The point-to-point error is normalized by the inter-ocular distance.*

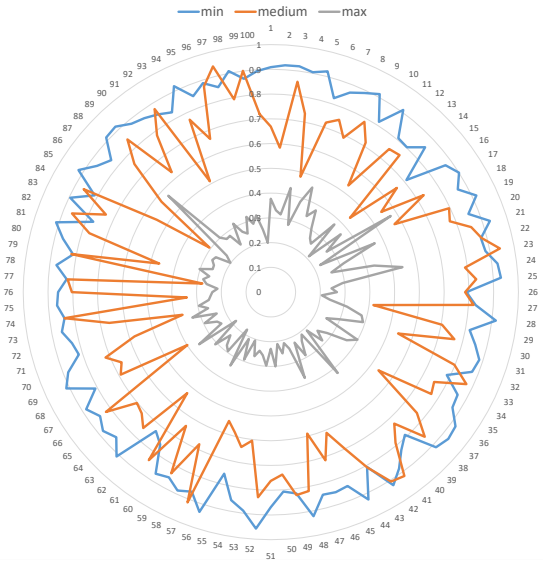


FIGURE 6: *Results of the discriminator network P for the task of facial landmark estimation. The samples are sorted from the highest NRMSE error to the lowest one. The discriminator scores of the top 100 samples are marked with the gray line. The medium 100 samples are marked with the orange line. The lowest 100 samples are marked with the blue line.*

LFPW and HELEN is commonly refereed as the *common test set of 300W competition [56]*, 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/.

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 initializations, we follow the popular routine [2], [43] to augment samples by (0.75-1.25) scaling and ±25° in-plane rotations generated from a uniform distribution. To reduce computation complexity, the network starts with a 7×7 convolutional layer with stride 2 to drop the resolution down from 256×256 to 128×128. Then the proposed PoseNet is connected to the 128 feature maps. As occlusion information is not provided in these datasets, we remove the occlusion sub-net in our framework. The networks is stacked four times in this task. For implementation, we train all our human pose models with the Torch7 toolbox [60].

3.1.1 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 which is consistent with [56]. Our method is compared to the state-of-the-art methods of Deep Alignment Network (DAN) [3], Mnemonic Descent Method [2], Coarse-to-Fine Shape Searching (CFSS) [61], Coarse-to-Fine Auto-encoder Networks (CFAN) [43], Local Binary Features (LBF) [40], Explicit Regression Trees (ERT) [62], Supervised Descent Method (SDM) [39], Explicit Shape Regression (ESR) [35],Deng *et.al* [63], Fan *et.al* [64], Martinez *et.al* [65], Uricar*et.al* [65], Face++ [66] and Yan *et.al* [67]. Results of the last six methods as listed are downloaded from the 300W competition website, SDM is implemented by ourself using the dense-SIFT feature provided by the author. For other methods, publicly available codes are used. The results show that PoseNet outperforms the rest of face alignments methods in every error metrics. It should be noted that although our method avoids the coarse-to-fine approaching strategy, we perform much better in fine estimation. Specially 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 achieve 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, as long as we give appropriate supervision during training.

We have calculated some further metrics from the CED curves to offer insight 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 original competition are shown in the table, as in Table 1. It can be shown although our method improves little in error rate when the threshold is set at 0.08, our method greatly reduces the mean error and improves the AUC performance a lot.

3.1.2 Quantitative Comparisons

To intuitively show the improvement of our method over previous methods, we show samples with large errors using previous methods in Fig 7. 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 successes to locate the landmarks without occlusions, it fails in the part of occluded mouth and surrounding face contour. Specially for the face contour, the landmarks are sorted discretely without any shape constrain. On the other side, our method successes in locating the landmarks and maintain the reasonable face shape.

To further show the usefulness of the discriminator network, we display the result scores in Fig. 6. 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 estimates are mostly extremely high, which will not help in observing the usefulness of P. Hence, we use a non-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 [41], [68], [69] for evaluation. We uniformly sample 4397 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. 7, it can be easily observed 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 ability of the designed discriminator. As long as the generator successfully "deceives" this discriminator, the landmark estimations become more accurate.

3.2 2D Human Pose Estimation

Datasets. We evaluate the proposed method on two widely used benchmarks on pose estimation, *i.e.*, extended Leeds Sports Poses (LSP) [70] and MPII Human Pose [71]. The LSP dataset consists of 11k training images and 1k testing images from sports activities. The MPII dataset consists of around 25k images with 40k annotated samples (about 28k for training, 11k for testing). The figures are annotated with 16 landmarks on the whole body with various challenging directions to the camera. On MPII, we train our model on a subset of training images while evaluating on the official test set and a held-out validation set about 3000 samples [5], [9]. Both datasets provide the visibility of body parts, which is used as the supervision occlusion signal in our method.

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'14 [35]	8.47	26.09	30.50
ERT'14 [62]	8.41	27.01	28.83
LBF'14[1] [40]	8.57	25.27	33.67
Yan'13 [67]	–	34.79	12.67
Face++'13 [66]	–	32.23	13.00
SDM'13 [39]	5.83	36.27	13.00
CFAN'14 [43]	5.78	34.78	14.00
CFSS'15 [61]	5.74	36.58	12.33
MDM'16 [2]	4.78	45.32	6.80
DAN'17 [3]	4.30	47.00	2.67
Ours	**3.96**	**53.64**	**2.50**

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

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 the training on 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 module and a max pooling to decrease 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).

3.2.1 Quantitative Results

We use the Percentage Correct Keypoints (PCK@0.2) [72] 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 [71] (referred to as PCKh@0.5).

LSP Human Pose. Table 2 shows the PCK performance of our method and previous methods at a normalized distance of 0.2. Our approach outperforms the state-of-the-art across all the body joints, and obtains improvement of 2.4% in average.

MPII Human Pose. Table 3 and Fig. 8 reports the PCKh performance of our method and previous methods at a normalized distance of 0.5. Baseline model refers to a

FIGURE 7: *Samples on the 300W test set. The four rows are results of MDM [2] , CFSS [61], SDM [39] 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.*

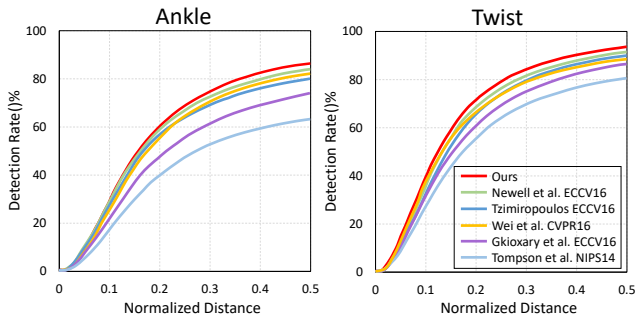


FIGURE 8: *PCKh comparison on the MPII validation set.*

TABLE 2: *Comparisons of PCK@0.2 performance on the LSP dataset.*

Methods	*Head*	*Sho.*	*Elb.*	*Wri.*	*Hip*	*Knee*	*Ank.*	**Mean**
B&Z'17 [73]	95.2	89.0	81.5	77.0	83.7	87.0	82.8	85.2
Lifshitz'16 [53]	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
Bulat'16 [10]	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Ours	**98.5**	**94.0**	**89.8**	**87.5**	**93.9**	**94.1**	**93.0**	**93.1**

TABLE 3: *Results on the 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'16 [74]	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
Hu'16 [75]	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 [53]	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Gkioxari'16 [76]	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Rafi'16 [77]	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
Bulat'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
Chu'17 [26]	**98.5**	96.3	91.9	88.1	**90.6**	88.0	85.0	91.5
Ours (test)[1]	98.1	**96.5**	**92.5**	**88.5**	90.2	**89.6**	**86.0**	**91.9**
Ours (-valid)[2]	98.2	96.2	90.9	86.7	89.8	87.0	83.2	90.6
Ours (valid)[3]	98.6	96.4	92.4	88.6	91.5	88.6	85.7	92.1

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

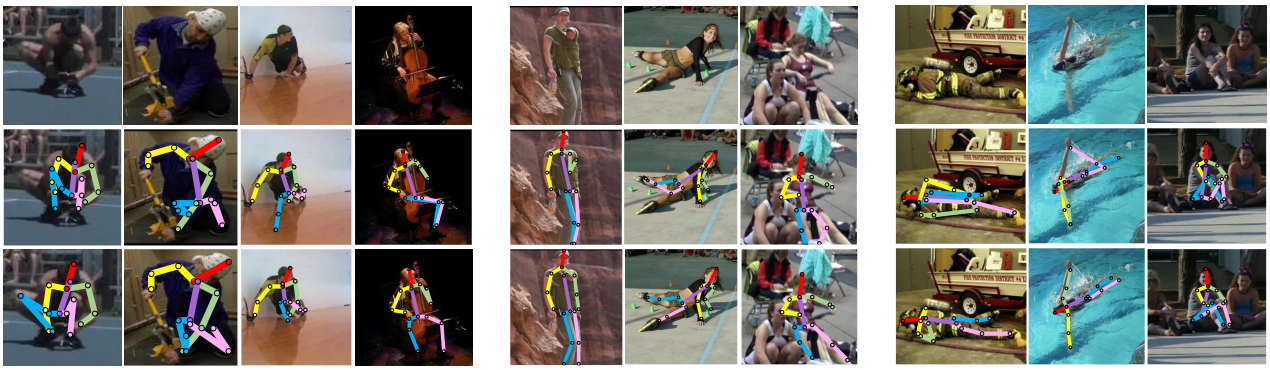
four-stacked single-task network without multi-task and discriminators. It has similar structure but half of stacked layers and parameter numbers compared to [9]. Our method achieves the best PCKh score of 91.9% on the test set.

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.

3.2.2 Quantitative Comparisons

To gain insights on how the proposed method accomplish the goal of setting the pose estimations within the geometric constraints, we visualize the predicted poses on the MPII test set compared with a 2-stacked hourglass network (HG) [9], as demonstrated in Fig. 9. For fair comparison, we also use a 2-stacked network in this section. We can see that our method gains a better understanding of the human body which leads to less weird predictions.

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



(a) (b) (c)

FIGURE 9: *Prediction samples on the MPII test set. The first row: original images. The second row: results by stacked hourglass network (HG) [9]. The third row: results by our method. (a)-(c) stand for three kinds of failure with HG.*

FIGURE 10: *Failure cases caused by body part at the edge (the first-second columns), overlapping people (the third column) and invisible limbs (the fourth column). The results on the first and second rows are generated by our method and HG [9], respectively.*

difficult situations.

On the other hand, we also show some failure examples of our method on the MPII test set in Fig. 10. As shown in Fig. 10, our method may fail in some challenging cases with twisted limbs at the edge, overlapping people and occluded body parts. In some cases, human may also fail to figure out the correct pose at a glance. Even when our method fails in this situations, it also achieves more reasonable poses compared to previous method. Previous method may generate some poses which violate human body structure as shown in the first row of Fig. 10. When the network fails to find high-confidence locations around the person, it shifts to the surrounding area where the local features matches the trained features best. Lacking of shape constraint finally results in these absurd poses.

3.2.3 Occlusion Analysis

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

First, two examples with some body parts occluded are given in Fig. 11. In the first sample, two legs of the person are totally occluded by the table. In the corresponding occlusion maps, the occluded part are well predicted. Despite of the occlusions, the pose heatmaps generated by our method are mostly clear and Gaussian centered. This results in high scores in both pose prediction and confidence evaluation despite of occlusions.

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

Next, we compare the performance of our method under occlusions with a stacked hourglass network [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 4. 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.*

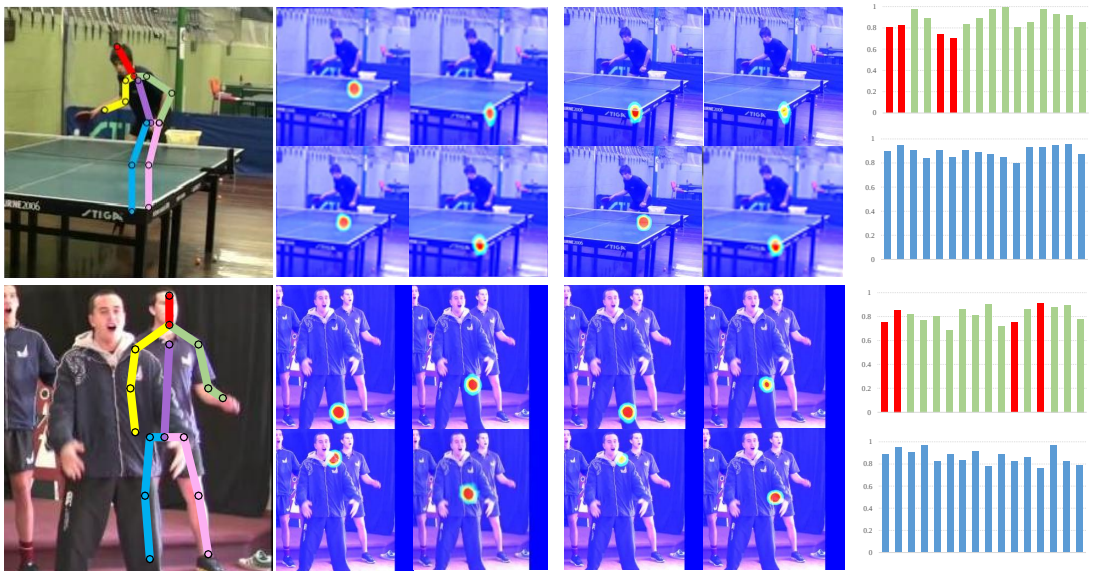
3.2.4 Ablation Study

To investigate the efficacy of the proposed multi-task generator network and the discriminators designed for learning human body priors, we conduct ablation experiments on the validation set of the MPII Human Pose dataset. A four-stacked single-task generator without occlusion is used as the baseline. The overall result is shown in Fig. 12. We give analysis to two components in our method: the multi-task manner and discriminators.

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

TABLE 4: *Detection rates (%) of visible and invisible elbows and wrists.*

Methods	Visible		Invisible	
	Wrist	*Elbow*	*Wrist*	*Elbow*
Newell'16 [9]	93.6	95.1	67.2	74.0
Ours	**94.5**	**95.9**	**70.7**	**77.6**



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

FIGURE 11: *(a) Input images with predicted poses; (b) Predicted pose heatmaps of four occluded body parts; (c) Predicted occlusion heatmaps of four occluded body parts; (d) Outputs values of P (in blue) and C(in green). Red columns in the output of C correspond to values of the four occluded body parts.*

TABLE 5: *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 the 3d baseline.*

Methods	*Direct.*	*Discuss*	*Eating*	*Greet*	*Phone*	*Photo*	*Pose*	*Purch.*	*Sitting*	*SitingD*	*Smoke*	*Wait*	*WalkD*	*Walk*	*WalkT*	**Mean**
LinKDE'14 [78] (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'15 [79] (MA)	–	136.9	96.9	124.7	–	168.7	–	–	–	–	–	–	132.2	70.0	–	–
Tekin'16 [80] (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
Hu'16 [75] (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'16 [30] (SA)	–	129.1	91.4	121.7	–	162.2	–	–	–	–	–	–	130.5	65.8	–	–
Ghezelghieh'16 [81] (SA)	80.3	80.4	78.1	89.7	–	–	–	–	–	–	–	–	–	95.1	82.2	–
Du'16 [82] (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'16 [83] (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'16 [28] (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'16 [31] (MA)	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
3d baseline [33] (SH,MA)	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) [33]	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

TABLE 6: *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'15 [84] (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'12 [85] (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'17 [86] (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'16 [29] (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'16 [32] (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
Pavlakos'16 [31] (SA)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	51.9
3d baseline [33] (SH,MA)	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
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**

Discriminator with Single-task. We also compare the four-stacked single-task generator trained with discriminators with the baseline. The networks are trained by removing the part for the occlusion heatmaps. Discriminators also

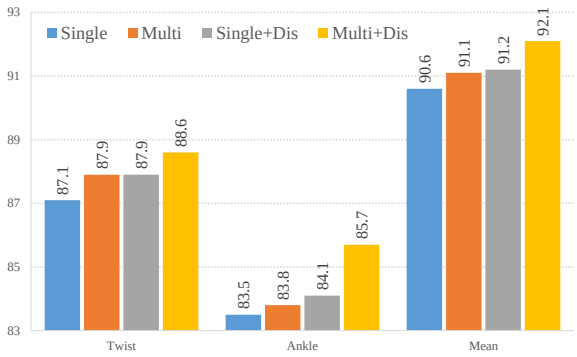


FIGURE 12: *Ablation study: PCKh scores at the threshold of 0.5.*

receive inputs without occlusion heatmaps. By using the body-structure-aware GANs, the performance on the MPII validation set increases by 0.6% compared to the baseline model. This shows that the discriminators contribute in pushing the generator to produce more reliable pose predictions.

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

3.3 2D to 3D Pose Transformation

Datasets and Experimental Settings We focus our numerical evaluation on a public dataset for 3d human pose estimation: Human3.6M [78]. Human3.6M 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 evaluation. For fair comparison to previous methods, we build our method based on a recently published baseline [33] and strictly follow their experimental settings.

In detail, the 2D to 3D transformation net is the same as [33]. 2D poses are estimated by the same hourglass networks as [33]. 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.* [29], [32]). We refer the experiment without further alignment as Protocol # 1 while the opposite 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 5 reports the results without further alignment and Table 6 reports the results with further alignment. By simply adding a structural PoseNet structure on [33], the performance is improved. It should be pointed out that this comes with no additional computation cost during test.

4 CONCLUSIONS

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

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

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

ACKNOWLEDGEMENTS

This work was in part supported by an ARC Future Fellowship to C. Shen; and an ARC DECTA Fellowship to L Liu. J. Yang's participation was supported by the National Science Fund of China under Grants #91420201, 61472187, 61502235, 61233011, 61373063 and 61602244, the 973 Program #2014CB349303, Program for Changjiang Scholars and Innovative Research Team in University.

Correspondence should be addressed to C. Shen.

REFERENCES

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

[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. Workshops of IEEE Conf. Comp. Vis. Patt. Recogn.*, 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," in *Proc. Int. Conf. Learn. Representations*, 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.*, 2017.

[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] 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.

[28] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei, "Deep kinematic pose regression," in *Proc. Workshops Eur. Conf. Comp. Vis.* Springer, 2016, pp. 186–201.

[29] 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.

[30] 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.

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

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

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] G. Tzimiropoulos, "Project-out cascaded regression with an application to face alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3659–3667.

[42] 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.

[43] 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.

[44] Y. Chen, W. Luo, and J. Yang, "Facial landmark detection via pose-induced auto-encoder networks," in *Proc. IEEE Int. Conf. Image Process.* IEEE, 2015, pp. 2115–2119.

[45] Y. Chen, J. Qian, J. Yang, and Z. Jin, "Face alignment with cascaded bidirectional lstm neural networks," in *Proc. Int. Conf. Patt. Recogn.* IEEE, 2016, pp. 313–318.

[46] M. Mirza and S. Osindero, "Conditional generative adversarial nets," in *Deep Learning Workshop, Advances in Neural Inf. Process. Syst.*, 2014.

[47] 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.

[48] 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.

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

[50] 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.

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

[52] 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.

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

[54] 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.

[55] 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.

[56] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. Workshops of IEEE Conf. Comp. Vis. Patt. Recogn.*, 2013, pp. 397–403.

[57] ——, "A semi-automatic methodology for facial landmark annotation," in *Proc. Workshops of IEEE Conf. Comp. Vis. Patt. Recogn.*, 2013, pp. 896–903.

[58] 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.

[59] 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.

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

[61] 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.

[62] 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.

[63] 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.

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

[65] 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.

[66] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *Proc. Workshops of IEEE Int. Conf. Comp. Vis.*, 2013, pp. 386–391.

[67] J. Yan, Z. Lei, D. Yi, and S. Li, "Learn to combine multiple hypotheses for accurate face alignment," in *Proc. Workshops of IEEE Int. Conf. Comp. Vis.*, 2013, pp. 392–396.

[68] G. G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape, "Offline deformable face tracking in arbitrary videos," in *Proc. Workshops of IEEE Int. Conf. Comp. Vis.*, 2015.

[69] 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 *Proc. Workshops of IEEE Int. Conf. Comp. Vis.*, 2015, pp. 50–58.

[70] 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.

[71] 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.

[72] 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.

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

[74] 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.

[75] 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.

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

[77] 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.

[78] 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.

[79] 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.

[80] 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.

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

[82] 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.

[83] S. Park, J. Hwang, and N. Kwak, "3d human pose estimation using convolutional neural networks with 2d pose information," in *Proc. Workshops Eur. Conf. Comp. Vis.* Springer, 2016, pp. 156–169.

[84] 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.

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

[86] 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.