# Adversarial Semantic Data Augmentation for Human Pose Estimation

Yanrui Bin[1[0000−0003−2845−3928]], Xuan Cao[2], Xinya Chen[1[0000−0002−6537−4316]], Yanhao Ge[2], Ying Tai[2], Chengjie Wang[2], Jilin Li[2], Feiyue Huang[2], Changxin Gao[1[0000−0003−2736−3920]], and Nong Sang[1[0000−0002−9167−1496]⋆]

[1] Key Laboratory of Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China
{yrbin, hust_cxy, cgao, nsang}@hust.edu.cn
[2] Tencent Youtu Lab
{marscao, halege, yingtai, jasoncjwang, jerolinli, garyhuang}@tencent.com

**Abstract.** Human pose estimation is the task of localizing body keypoints from still images. The state-of-the-art methods suffer from insufficient examples of challenging cases such as symmetric appearance, heavy occlusion and nearby person. To enlarge the amounts of challenging cases, previous methods augmented images by cropping and pasting image patches with weak semantics, which leads to unrealistic appearance and limited diversity. We instead propose Semantic Data Augmentation (SDA), a method that augments images by pasting segmented body parts with various semantic granularity. Furthermore, we propose Adversarial Semantic Data Augmentation (ASDA), which exploits a generative network to dynamiclly predict tailored pasting configuration. Given off-the-shelf pose estimation network as discriminator, the generator seeks the most confusing transformation to increase the loss of the discriminator while the discriminator takes the generated sample as input and learns from it. The whole pipeline is optimized in an adversarial manner. State-of-the-art results are achieved on challenging benchmarks. The code has been publicly available at https://github.com/Binyr/ASDA.

**Keywords:** Pose Estimation, Semantic Data Augmentation

## 1 Introduction

Human Pose Estimation (HPE) is the task of localizing body keypoint from still images. It serves as a fundamental technique for numerous computer vision applications. Recently, deep convolutional neural networks (DCNN) [23,13,33] have achieved drastic improvements on standard benchmark datasets. However, as shown in Figure 1, they are still prone to fail in some challenging cases such as symmetric appearance, heavy occlusion, and nearby persons.

⋆ Corresponding author.

(a) Symmetric Appearance (b) Occlusion (c) Nearby Person

**Fig. 1.** Pairs of pose predictions obtained by HRNet [23] (top) and our approach (bottom) in the challenging cases. Incorrect predictions are highlighted by the red dotted circles. Note that image in Figure 1 (c) {*cols. 1*} is an extremely challenging case so that few of the keyponts are correctly predicted by the original HRNet. After equipped with our ASDA (bottom), HRNet improve the robustness to the challenging cases.

The reason for the inferior performance of the DCNN-based methods in the challenging cases is that there exists an insufficient amount of examples that contain these challenging cases to train a deep network for accurate keypoint localization. However, obtaining the annotations of keypoint localization is costly.

One promising way to tackle this problem is data augmentation. Conventional data augmentation performs global image transformations (e.g., scaling, rotating, flipping or color jittering). Although it enhances the global translational invariance of the network and largely improves the generalizability, it contributes little to solving the challenging cases. Recently, Ke et al. [13] proposes keypoints masking training to force the network better recognize poses from difficult training samples. They simulate the keypoint occlusion by copying a background patch and putting it onto a keypoint or simulate the multiple existing keypoints by copying a body keypoint patch and putting it onto a nearby background. However, this data augmentation method only brings marginal improvement. On the one hand, the used patch is cropped from the input image, resulting in a limited variance of the generated images. On the other hand, the cropped keypoint patch is surrounded by some background, which makes the generated image unrealistic.

In this paper, we propose a novel Adversarial Semantic Data Augmentation (ASDA) scheme. Human parsing is applied to the training images to get a large amount of pure body part patches. These body parts are organized, according to their semantic types, to build a semantic part pool. As the human body could be represented as a hierarchy of parts and subparts, we combine several subparts, according to the structure of the human body, to get body parts with various semantic granularity. For each input image, several parts will be randomly selected from the semantic part pool and properly pasted to the image.

Further, randomly pasting parts to the image is still suboptimal. Without taking the difference between training image instances into account, it may generate ineffective examples that are too easy to boost the network. Moreover, it can hardly match the dynamic training status of the pose estimation network, since it is usually sampled from static distributions [21]. For instance, with the training of the network, it may gradually learn to associate occluded wrists while still have difficulty in distinguish similar appearance with legs.

Based on the above consideration, we parameterize the parts pasting process as an affine transformation matrix and exploit a generative network to online predict the transformation parameters. The generator seeks the most confusing transformation to increase the loss of the pose estimation network and consequently generates tailored training samples. The pose estimation network acts as a discriminator, which takes the tailored samples as input and tries to learn from it. By leveraging the spatial transformer network, the whole process is differentiable and trained in an adversarial manner.

Additionally, our Adversarial Semantic Data Augmentation is a universal solution that can be easily applied to different datasets and networks for human pose estimation.

In summary, the main contributions are three-fold:

• We design a novel Semantic Data Augmentation (SDA) which augments images by pasting segmented body parts of various semantic granularity to simulate examples that contain challenging cases.

• We propose to utilize a generative network to dynamically adjust the augmentation parameters of the SDA and produce tailored training samples against the pose estimation network, which largely elevates the performance of the SDA.

• We comprehensively evaluate our methods on various benchmark datasets and consistently outperforms the state-of-the-art methods.

## 2 Related Work

The advances of DCNN-based human pose estimation benefit from multiple factors. We compare our methods with literature from three most related aspects.

### 2.1 Human Pose Estimation.

Recently, pose estimation using DCNNs has shown superior performance. DeepPose [27] first applied deep neural networks to human pose estimation by directly regressing the 2D coordinates of keypoints from the input image. [26] proposed a heatmap representation for each keypoint and largely improved the spatial generalization. Following the heatmap-based framework, various methods [29,18,22,24,23,30,23] focused on designing the structure of the network and indeed achieved significant improvement. However they still suffered from insufficient amounts of samples that contains challenging cases. In this work, standing on the shoulder of the well-designed network structure, we propose a universal data augmentation solution to further improve the performance of human pose estimation.

### 2.2 Data Augmentation.

Typical data augmentation [18,4,30,23] mainly performed global spatial transformation like scaling, rotating and flipping *etc.* These common data augmentation schemes helped the network to resist the global image deformation but fail to improve the immunity to the challenging cases. Recently, some novel data augmentations were proposed. PoseRefiner [8] transformed the keypoint annotations to mimic the most common failure cases of human pose estimators, so that the proposed refiner network could be trained well. MSR-net [13] introduced keypoint-masking which cropped and pasted patches from the input image to simulate challenging cases. Different from the existing data augmentation strategies, we propose a novel semantic data augmentation scheme which takes advantage of the human semantic segmentation to obtain the pure segmented body parts rather than noisy image patches. Furthermore, we compose the related parts to form a set of new parts with higher semantic granularity.

### 2.3 Adversarial Learning.

Inspired by the minimax mechanism of Generative Adversarial Networks (GANs) [10], some literature [5] generated hard training samples in an adversarial way. Semantic Jitter [32] proposed to overcome the sparsity of supervision problem via synthetically generated images. A-Fast-RCNN [28] used GANs to generate deformations for object detection. Recently, GANs were introduced into human pose estimation. Such as Adversarial PoseNet [4] designed discriminators to distinguish the real poses from the fake ones. Jointly Optimize [21] designed an augmentation network that competed against a target network by generating hard augmentation operations. In this paper, we designed a generative network to adjust the semantic data augmentation then to produce challenging training data. The generative network takes the difference between training instances into consideration, and produce tailored training samples for the pose estimation network. Hard mining, as an alternative strategy to feed challenging training data to network, is totally different from ours. Hard mining can only "select" rather than "produce" challenging samples, which essentially limits its improvement of accuracy on challenging cases.

## 3 Methodology

### 3.1 Semantic Data Augmentation

**Building Semantic Part Pool.** For common human pose estimation schemes [18,30,25,23], data augmentations such as global scaling, rotation, flipping are usually applied, which bring the global translational invariances to the network and largely improves the generalizability.

However, the remained problem of pose estimation task is the challenging cases, e.g., symmetric appearance, heavy occlusion, and nearby person, where the global spatial transformation helps little. In contrast to the global spatial


**Fig. 2.** Illustration of Semantic Data Augmentation (SDA). We first apply human parsing on training images and get a large amount of segmented body parts. The segmented body parts are organized, according to their semantics, to build semantic part pool. For each training image, several part patches will be randomly sampled and properly placed on the image to synthesize the real challenging cases such as symmetric appearance (green circle), occlusion (perple circle) and nearby person (yellow circle).

transformations, local pixel patch manipulation provide more degrees of freedom to augment image and is able to synthesize the challenging case realistically.

A human image is assembled by semantic part patches, such as arm, leg, shoe, trousers and so on. Inspired by these semantic cues, we can synthesize plentiful human image instances by elaborately combining these local part patches. Here, we propose a novel augmentation scheme, as shown in Figure 2. By firstly segmenting all human images through the human parsing method [17], then we can build a data pool $\mathbb{D}_{part}$ filled with various semantic body part patches. We follow the definition of LIP dataset [9] and segment the human image into $\hat{N} = 26$ part patches. Finally, the body part patches from the data pool can be properly mounted on the current person's body to synthesize challenging cases.

As human parsing aims to analyze every detail region of a person as well as different categories of clothes, LIP defines 6 body parts and 13 clothes categories in fine semantic granularity. However, body parts of various semantic granularity will appear in images of real-world scenarios with complex multi-person activities. For the above considerations, we combine some of the parts (e.g., left shoe and left leg) to form a set of new parts with higher semantic granularity and then add them to our part pool. After the cutting step, we filter out scattered segments, segments with the area below $35^2$ and segments with low semantics.

**Augmentation Parameter Formulation.** Given a semantic part patch $I_p$ and a training image $I_o$, the placement of this semantic part can be defined by the affine transformation matrix

$$\boldsymbol{H} = \begin{bmatrix} s\cos r & s\sin r & t_x \\ -s\sin r & s\cos r & t_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{1}$$

where $s$ denotes the scale of the part patch, $r$ denotes the rotation, and $t_x, t_y$ is the translation in horizontal and vertical direction respectively. Thus the placement of the part patch $I_p$ can be uniquely determined by a 4D tuple $\theta(s, r, t_x, t_y)$.

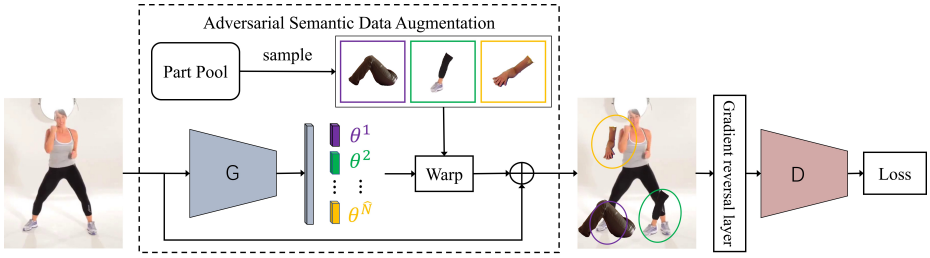


**Fig. 3.** Overview of our approach. The input image is fed to the generator $\mathcal{G}$ to obtain $\hat{N}$ groups of tailored augmentation parameters which are used to warp the randomly selected semantic part patches. Each group parameters is used to warp the patch of the specific part type. $\mathcal{G}$ seeks the most confusing transformation to increase the loss of the pose estimation network and consequently generates tailored training samples. The pose estimation network acts as a discriminator $\mathcal{D}$, which takes the tailored sample as input and tries to learn from it. The whole pipeline is optimized in an adversarial manner.

The scale of the part patch will be aligned with the target person in advance according to the human bounding box. Initially, the part patch could be pasted in the center of the training image without rotation. In other words, the tuple $(1, 0, 0, 0)$ is served as our original paste configuration.

**Random Semantic Augmentation.** With 4D augmentation parameters defined in Equation 1, a straight augmentation method can be realized by sampling a 4D tuple augmentation parameter from a uniform distribution in the neighborhood space of $(1, 0, 0, 0)$. $N$ different body parts will be pasted to the target person. The value of $N$ is set manually as a hyper-parameter. Sensitivity Analysis of $N$ is detailed in Section 4.5.

### 3.2 Adversarial Learning

Our goal is to generate the confusing transformation to improve the performance of pose estimation networks. However, the augmentation parameters of SDA are sampled from the neighborhood of $(1, 0, 0, 0)$. On the one hand, the confusing transformation naturally varies with different training instances and different part types. On the other hand, random sampling augmentation parameters from the static distribution can hardly perceive the dynamic training status. Thus it is prone to generate ineffective training samples which are so easy that it may not bring positive or even put negative effect on network training.

To overcome such issues, we propose to leverage Spatial Transformer Network (STN) to manipulate semantic parts within the network and optimize it in an adversarial manner. The main idea is to utilize an STN as the generator, which seeks the most confusing transformation to increase the pose estimation network loss. On the other hand, the pose estimation network acts as a discriminator, which tries to learn from the tailored semantic augmentation.

**Generate Tailored Samples.** The core module of our method is an STN, which takes the target person image as input and predicts $\hat{N}$ groups transformation parameters, each of which is used to transform the randomly selected semantic body parts of the specific part type. In our experiments, we find that allowing the network to predict the scale $s$ of the part would collapse the training. It would easily predict a large scale, so that the part completely covers the target person in the training images. Thus, we randomly sample the scale $s$ from the neighboring space of 1.0 and the generative network is mainly responsible for predicting the $(r, t_x, t_y)$. The affine transformation matrix is generated as defined in Equation 1.

Each pixel in the transformed image is computed by applying a sampling kernel centered at a particular location in the original image. Mathematically, the pointwise transformation is shown in eq. (2).

$$\begin{pmatrix} x_i^s \\ y_i^s \\ 1 \end{pmatrix} = \boldsymbol{H} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} , \tag{2}$$

where $(x_i^s, y_i^s)$ and $(x_i^t, y_i^t)$ denote the coordinates of the i-th pixel in the original and transformed image respectively. The transformed parts thus can be pasted to the target person image in the order they were sampled.

It is the not the first time to determine the augmentation parameters through a network. Xi Peng et al [21] jointly optimizes the conventional data augmentation (i.e., global scaling, rotating and feature erasing.) and network training to enhance the global transformation invariance of the network. Our contributions are quite different with [21]. We design a novel SDA which augments images by pasting segmented body parts of various semantic granularity to simulate examples that contain challenging cases. Then we further propose ASDA that utilize a generative network to dynamically adjust the augmentation parameters of the SDA and produce tailored training samples for the pose estimation network.

**Joint Training.** As shown in the Figure 3, the networks training follow the pipeline of training standard GANs [10]. Generative network acting as generator $\mathcal{G}$ try to produce challenging cases. Meanwhile, the pose estimation network acting as a discriminator $\mathcal{D}$ try to learn from the generated training samples.

The discriminator is supervised by ground-truth heatmaps and try to decrease the loss $\mathcal{L}_{\mathcal{D}}$ which is formulated as eq. (4). On the contrary, the generator try to increase the loss $\mathcal{L}_{\mathcal{D}}$. So the loss for generator is simply set as negative discriminator loss as formulated in eq. (5).

$$I_{aug} = \mathcal{F}_{aff}(\mathcal{G}(I_o), \{I_p\}) , \tag{3}$$

$$\mathcal{L}_{\mathcal{D}} = \|\mathcal{D}(I_{aug}) - H_{gt}\|_{\ell_2} , \tag{4}$$

$$\mathcal{L}_{\mathcal{G}} = -\mathcal{L}_{\mathcal{D}} , \tag{5}$$

where $I_o$ is the original training image, $\{I_p\}$ is a set of randomly sampled part patches, $\mathcal{F}_{aff}(\cdot, \cdot)$ denotes the affine transformation function and $H_{gt}$ denote ground-truth heatmap. The network weights of $\mathcal{G}$ and $\mathcal{D}$ are updated alternately.

# 4 Experiments

## 4.1 Datasets and Evaluation Protocols

We conduct experiments on three representative benchmark datasets, *i.e.* extended Leeds Sports Poses (LSP) dataset [12], MPII human pose dataset [1] and MS COCO dataset [16].

**LSP Dataset.** The extended LSP dataset consists of 11k training images and 1k testing images of mostly sports people. Standard Percentage of Correct Keypoints (PCK) metric is used for evaluation. It reports the percentage of keypoint that fall into a normalized distance of the ground-truth, where the torso size is used as the normalized distance.

**MPII Dataset.** The MPII dataset includes around 25k images containing over 40k people with annotated body keypoint (28k training and 11k testing). Following [18], 3k samples are taken as a validation set to tune the hyper-parameters. PCK is also utilized to evaluate MPII, but distance is normalized by head size. MPII evaluation metric is referred to PCKh.

**COCO Dataset.** The COCO dataset involves multi-person pose estimation task which requires simultaneously detecting people and localizing their key points. The COCO training dataset (train2017) includes 57k images and validation dataset (val2017) includes 5000 images. The COCO evaluation defines the object keypoint similarity (OKS) which plays the same role as the IoU.

## 4.2 Implementation Details

Both generator $\mathcal{G}$ and discriminator $\mathcal{D}$ are the off-the-shelf networks. For generator, the ResNet-18 is utilized to regress $(3 \times \hat{N})$ parameters, where $\hat{N}$ is the class number of the human parsing. For discriminator, we adopt HRNet [23].

During building the semantic part pool, in order to avoid the inference of different human parsing algorithms, we obtain body parts from LIP dataset [9]. Beside our semantic data augmentation, we keep original data augmentation as adopted in HRNet, including global random flip, rotation and scale.

Network training is implemented on the open-platform PyTorch. For training details, we employ Adam [14] with a learning rate 0.001 as the optimizer of both generator and discriminator network. We drop the learning rate by a factor of 10 at the 170-th and 200-th epochs. Training ends at 210 epochs. The HRNet is initialized with weight of pre-trained model on public-released ImageNet [7].

**MPII.** For both MPII training and testing set, body scale and center are provided. We first utilize these value to crop the image around the target person and resized to $256 \times 256$ or $384 \times 384$. Data augmentation includes random flip, random rotation $(-30°, 30°)$ and random scale $(0.75, 1.25)$.

**LSP.** For LSP training set, we crop image by estimating the body scale and position according to keypoint positions. The data augmentation strategy are the same to MPII. For the LSP testing set, we perform similar cropping and resizing, but simply use the image center as the body position, and estimate the body scale by the image size following [31]. We follow previous methods [29,31] to

train our model by adding the MPII training set to the extended LSP training set with person-centric annotations. For both MPII and LSP, testing is conducted on six-scale image pyramids (0.8, 0.9, 1.0, 1.1, 1.2 1.3).

**COCO.** For COCO training set, each ground-truth human box is extended to fixed aspect ratio, e.g., height : width = 4 : 3 and enlarged to contain more context by a rescale factor 1.25. Then the resulting box is cropped from image without distorting image aspect ratio and resized to a fixed resolution. The default resolution is 256 : 192. We apply random flip, random rotation $(-40°, 40°)$ and random scale $(0.7, 1.3)$. For COCO testing set, we utilized the predicted bounding box released by Li et al [15]. We also predict the pose of the corresponding flipped image and average the heat maps to get the final prediction.

### 4.3 Quantitative Results

We report the performance of our methods on the three benchmark datasets following the public evaluation protocols. We adopt the HRNet as the backbone network. "W32" and "W48" represent the channel dimentions of the high-resolution subnetworks in last three stages of HRNet, respectively. "s7" indicates the we expand the HRNet to 7 stages by repeating the last stage of the original HRNet.

**Results on LSP.** Table 1 presents the PCK@0.2 scores on LSP test set. Our method outperforms the state-of-the-art methods especially on some challenging keypoints, e.g., wrist, knee and ankle, we have 0.8%, 1.0% and 1.0% improvements respectively.

**Table 1.** Comparisons on the LSP test set (PCK@0.2).

| Method | Hea. | Sho. | Elb. | Wri. | Hip. | Kne. | Ank. | Total |
|---|---|---|---|---|---|---|---|---|
| Insafutdinov et al., 2016 [11] | 97.4 | 92.7 | 87.5 | 84.4 | 91.5 | 89.9 | 87.2 | 90.1 |
| Wei et al., 2016 [29] | 97.8 | 92.5 | 87.0 | 83.9 | 91.5 | 90.8 | 89.9 | 90.5 |
| Bulat et al., 2016 [2] | 97.2 | 92.1 | 88.1 | 85.2 | 92.2 | 91.4 | 88.7 | 90.7 |
| Chu et al., 2017 [6] | 98.1 | 93.7 | 89.3 | 86.9 | 93.4 | 94.0 | 92.5 | 92.6 |
| Chen et al., 2017 [4] | 98.5 | 94.0 | 89.8 | 87.5 | 93.9 | 94.1 | 93.0 | 93.1 |
| Yang et al., 2017 [31] | 98.3 | 94.5 | 92.2 | 88.9 | 94.4 | 95.0 | 93.7 | 93.9 |
| Zhang et al., 2019 [33] | 98.4 | 94.8 | 92.0 | 89.4 | 94.4 | 94.8 | 93.8 | 94.0 |
| Ours-W32 | **98.8** | **95.2** | **92.5** | **90.2** | **94.7** | **95.8** | **94.8** | **94.6** |

**Results on MPII.** The performance of our methods on MPII test set is shown in Table 2. We can observe that Ours-W48-s7 achieves 94.1% PCKh@0.5, which is the new state-of-the-art result. In particular, Ours-W48-s7 achieves 0.5%, 0.5% and 0.7% improvements on wrist, knee and ankle which are considered as the most challenging keypoints.

**Results on COCO.** Table 3 compares our methods with classic and SOTA methods on COCO val2017 dataset. All the methods use standard top-down paradigm which sequentially performs human detection and single-person pose

**Table 2.** Comparisons on the MPII test set (PCKh@0.5).

| Method | Hea. | Sho. | Elb. | Wri. | Hip. | Kne. | Ank. | Total |
|---|---|---|---|---|---|---|---|---|
| Wei et al., 2016 [29] | 97.8 | 95.0 | 88.7 | 84.0 | 88.4 | 82.8 | 79.4 | 88.5 |
| Bulat et al., 2016 [2] | 97.9 | 95.1 | 89.9 | 85.3 | 89.4 | 85.7 | 81.7 | 89.7 |
| Newell et al., 2016 [18] | 98.2 | 96.3 | 91.2 | 87.1 | 90.1 | 87.4 | 83.6 | 90.9 |
| Ning et al., 2018 [20] | 98.1 | 96.3 | 92.2 | 87.8 | 90.6 | 87.6 | 82.7 | 91.2 |
| Chu et al., 2017 [6] | 98.5 | 96.3 | 91.9 | 88.1 | 90.6 | 88.0 | 85.0 | 91.5 |
| Chen et al., 2017 [4] | 98.1 | 96.5 | 92.5 | 88.5 | 90.2 | 89.6 | 86.0 | 91.9 |
| Yang et al., 2017 [31] | 98.5 | 96.7 | 92.5 | 88.7 | 91.1 | 88.6 | 86.0 | 92.0 |
| Xiao et al., 2018 [30] | 98.5 | 96.6 | 91.9 | 87.6 | 91.1 | 88.1 | 84.1 | 91.5 |
| Ke et al., 2018 [13] | 98.5 | 96.8 | 92.7 | 88.4 | 90.6 | 89.4 | 86.3 | 92.1 |
| Nie et al., 2018 [19] | 98.6 | 96.9 | 93.0 | 89.1 | 91.7 | 89.0 | 86.2 | 92.4 |
| Tang et al., 2018 [25] | 98.4 | 96.9 | 92.6 | 88.7 | 91.8 | 89.4 | 86.2 | 92.3 |
| Sun et al., 2019 [23] | 98.6 | 96.9 | 92.8 | 89.0 | 91.5 | 89.0 | 85.7 | 92.3 |
| Zhang et al., 2019 [33] | 98.6 | 97.0 | 92.8 | 88.8 | 91.7 | 89.8 | 86.6 | 92.5 |
| Su et al., 2019 [22]* | 98.7 | 97.5 | 94.3 | 90.7 | **93.4** | 92.2 | 88.4 | 93.9 |
| Ours-W48-s7* | **98.9** | **97.6** | **94.6** | **91.2** | 93.1 | **92.7** | **89.1** | **94.1** |

"*" indicates the network take image size $384 \times 384$ as input.

estimation. Our model outperforms SIM [30] and HRNet [23] by 4.8% and 0.8% for input size $256 \times 192$ respectively. When input size is $384 \times 288$, our model achieve better AP than SIM [30] and HRNet [23] by 4.5% and 0.9%.

**Table 3.** Comparison with SOTA methods on COCO val2017 dataset. Their results are cited from Chen et al. [3] and Sun et al. [23].

| Method | Backbone | Input Size | Params | GFLOPs | AP | $AP^{50}$ | $AP^{75}$ | $AP^{M}$ | $AP^{L}$ | AR |
|---|---|---|---|---|---|---|---|---|---|---|
| Hourglass [18] | HG-8stage | $256 \times 192$ | 25.1M | 14.3 | 66.9 | - | - | - | - | - |
| CPN [3] | ResNet-50 | $256 \times 192$ | 27.0M | 6.20 | 69.4 | - | - | - | - | - |
| CPN [3] | ResNet-50 | $384 \times 288$ | 27.0M | 13.9 | 71.6 | - | - | - | - | - |
| SIM [30] | ResNet-50 | $256 \times 192$ | 34.0M | 8.9 | 70.4 | 88.6 | 78.3 | 67.1 | 77.2 | 76.3 |
| SIM [30] | ResNet-50 | $384 \times 288$ | 34.0M | 20.0 | 72.2 | 89.3 | 78.9 | 68.1 | 79.7 | 77.6 |
| HRNet [23] | HRNet-W32 | $256 \times 192$ | 28.5M | 7.10 | 74.4 | 90.5 | 81.9 | 70.8 | 81.0 | 79.8 |
| HRNet [23] | HRNet-W32 | $384 \times 288$ | 28.5M | 16.0 | 75.8 | 90.6 | 82.7 | 71.9 | 82.8 | 81.0 |
| Ours | HRNet-W32 | $256 \times 192$ | 28.5M | 7.10 | **75.2** | **91.0** | **82.4** | **72.2** | **81.3** | **80.4** |
| Ours | HRNet-W32 | $384 \times 288$ | 28.5M | 16.0 | **76.7** | **91.2** | **83.5** | **73.2** | **83.4** | **81.5** |

### 4.4 Qualitative Results

Figure 4 displays some pose estimation results obtained by HRNet without (left size) and with (right side) our ASDA. We can observe that original HRNet is confused by symmetric appearance (e.g. the left and right legs in {*rows.1, cols. 3*}), heavy occlusion (e.g., the right ankle in {*rows.1 cols. 2*}) and nearby person (e.g., multiple similar legs and arms in {*rows.1, cols. 1*}). Note that image

**Fig. 4.** Comparisons of the HRNet [23] trained without (left side) and with (right side) our Adversarial Semantic Data Augmentation.

in {*rows.1, cols. 1*} is an extremely challenging case so that few of the keyponts are correctly predicted by the original HRNet. By generating tailored semantic augmentation for each input image, our ASDA largely improves the performance of the original HRNet in the extremely challenging cases. Figure 5 shows some pose estimation results obtained by our approach on the COCO test dataset.

### 4.5 Ablation Studies

In this section, we conduct ablative analysis on the validation set of MPII dataset. The baseline is HRNet-W32 [23] which achieved PCKh@0.5 at 90.3% by performing flipping and single scale in inference. During baseline training, the data augmentation adopts global spatial transformation including random rotation $(30°, 30°)$ ,random scale $(0.75, 1.25)$ and flipping. The results are shown in Table 4 (a).

The MPII dataset provide visibility annotations for each keypoint, which enables us to conduct ablative analysis on the subset of invisible keypoints and study the effect of our method on improving the occlusion cases. The results are shown in Table 4 (b).

**With Vs. Without Semantic Data Augmentation.** We first evaluate the effect of the Semantic Data Augmentation scheme. As shown in Table 4 (a), **+SDA** outperforms the **Baseline** with a large margin by 0.5%. Note that our SDA scheme consistently achieved improvements on all keypoints. Especially, our SDA achieves 0.9%, 0.5% and 0.4% improvements on elbow, wrist and ankle respectly, which are considered as the most challenging keypoints to be localized. In Table 4 (b), we can observe a more significant improvement brought by SDA. The result demonstrate that the semantic local pixel manipulation of our SDA effectively augment training data and elevate the performance of pose estimation.

**Table 4.** Ablation studies on the MPII validation set (PCKh@0.5)

(a) Results evaluated on all keypoints

| Method | Hea. | Sho. | Elb. | Wri. | Hip. | Kne. | Ank. | Total |
|---|---|---|---|---|---|---|---|---|
| **Baseline** | 97.1 | 95.9 | 90.3 | 86.4 | 89.1 | 87.1 | 83.3 | 90.3 |
| **+ROR** | 97.0 | 96.2 | 90.9 | 86.9 | 89.3 | 86.9 | 82.9 | 90.5 |
| **+SDA (Ours)** | 97.2 | 96.3 | 91.2 | 86.9 | 90.0 | 87.2 | 83.7 | 90.8 |
| **+ASDA (Ours)** | **97.6** | **96.6** | **91.5** | **87.3** | **90.5** | **87.5** | **84.5** | **91.2** |

(b) Results evaluated only on invisible keypoints

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Baseline** | - | 90.9 | 73.6 | 61.9 | 81.8 | 71.7 | 61.8 | 74.2 |
| **+ROR** | - | 92.0 | 74.9 | 63.2 | 82.7 | 71.6 | 61.6 | 74.9 |
| **+SDA (Ours)** | - | 91.8 | 75.1 | 63.0 | 84.1 | 71.7 | 63.3 | 75.4 |
| **+ASDA (Ours)** | - | **92.7** | **75.1** | **65.1** | **84.8** | **71.8** | **63.4** | **76.1** |

**Baseline:** The original HRNet-W32 [23]. The following experiments is all based on this baseline.
**+ROR:** Adopt data augmentation of Randomly Occluding and Repeating (ROR) the keypoints patch [13] on training HRNet-W32.
**+SDA:** Adopt our Semantic Data Augmentation (SDA) scheme on training HRNet-W32, the augmentation parameters are adjusted randomly from a uniform distribution in the neighborhood space of $(1, 0, 0, 0)$.
**+ASDA:** Adop our Adversarial Semantic Data Augmentation (ASDA) scheme on training HRNet-W32, the augmentation parameters are online adjusted by the generative network in an adversarial way.

Both SDA and Randomly Occluding and Repeating (ROR) the keypoints patch [13] augment training data by manipulate the local pixel. However, ROR achieves 0.3% lower average PCKh@0.5 than our SDA. Moreover, ROR even brings negtive effects to baseline model when localizing keypoints like knee and ankle. These results demonstrate that various segmented body parts with high semantics used in our SDA play an key role for improving pose estimtion performance.

**Random Vs. Adversarial Augmentation.** Based on the SDA scheme, we found that Adversarial SDA can further improve the accuracy by online adjusting augmentation parameters. As shown in the table 4 (a), **+ASDA** consistently outperforms **+SDA** on all keypoints and achieve 0.4% higher average PCKh@0.5. For invisible keypoints, ASDA outperforms baseline and SDA by 1.9% and 0.7% PCKh@0.5 score. As discussed in Sec. 3.2, our ASDA can further improve performance due to the adversarial learning strategy which generates tailored samples for training pose estimation network.

**Sensitivity Analysis.** The part number $N$ as a hyper-parameter is configured manually. We test different $N$ values during training and the PCKh@0.5 score on the MPII validation set is shown in Table 5. Less than 3 parts, the performance maintain roughly the same. Begin with 4 parts, the performance sharply drop along the increasing of part number. We infer that too many parts

will generate too hard training samples for pose estimation network which misleads network to learn unrealistic cases.

**Table 5.** Ablation studies of different number of body parts $N$.

| Part Num | Hea. | Sho. | Elb. | Wri. | Hip. | Kne. | Ank. | Total |
|---|---|---|---|---|---|---|---|---|
| 1 | **97.6** | 96.6 | **91.5** | **87.3** | 90.5 | **87.5** | **84.5** | **91.2** |
| 2 | 97.5 | 96.6 | **91.5** | 86.9 | 90.1 | 87.4 | 83.8 | 91.0 |
| 3 | 97.3 | **96.8** | 91.3 | 86.9 | **90.6** | 87.4 | 83.6 | 91.0 |
| 4 | 97.4 | 96.3 | 91.1 | 86.2 | 90.3 | 87.0 | 83.6 | 90.7 |
| 6 | 97.2 | 96.2 | 90.4 | 85.2 | 90.0 | 86.0 | 82.1 | 90.1 |
| 8 | 97.0 | 95.7 | 89.3 | 83.8 | 89.3 | 85.6 | 81.4 | 89.4 |

**Apply on Different Networks.** As shown in Table 6, we report the performance of different networks trained with our ASDA. By applying our ASDA, the SOTA networks consistently achieved improvements. Especially on the challenging keypoints such as elbow, wrist, knee and ankle, our ADSA enhances the network significantly. This result exhibits the universality of our ADSA scheme.

**Table 6.** Result of applying on different network.

| Method | Hea. | Sho. | Elb. | Wri. | Hip. | Kne. | Ank. | Total |
|---|---|---|---|---|---|---|---|---|
| 2-Stacked HG | 96.6 | 95.4 | 89.7 | 84.7 | 88.7 | 84.1 | 80.7 | 89.1 |
| 2-Stacked HG+ASDA | **96.8** | **95.8** | **90.5** | **85.5** | **89.3** | **85.5** | **81.9** | **89.8** |
| 8-Stacked HG | 96.9 | 95.9 | 90.8 | 86.0 | 89.5 | 86.5 | 82.9 | 90.2 |
| 8-Stacked HG+ASDA | **97.5** | **96.5** | **91.6** | **87.3** | **90.5** | **87.7** | **83.5** | **91.1** |
| SIM-ResNet50 | 96.4 | 95.3 | 89.0 | 83.2 | 88.4 | 84.0 | 79.6 | 88.5 |
| SIM-ResNet50+ASDA | **96.8** | **95.8** | **89.7** | **83.9** | **89.5** | **85.1** | **80.5** | **89.3** |
| SIM-ResNet101 | 96.9 | **95.9** | 89.5 | 84.4 | 88.4 | 84.5 | 80.7 | 89.1 |
| SIM-ResNet101+ASDA | **97.2** | **95.9** | **90.0** | **85.2** | **89.7** | **86.0** | **82.3** | **90.0** |
| HRNet-W32 | 97.1 | 95.9 | 90.3 | 86.4 | 89.1 | 87.1 | 83.3 | 90.3 |
| HRNet-W32+ASDA | **97.6** | **96.6** | **91.5** | **87.3** | **90.5** | **87.5** | **84.5** | **91.2** |
| HRNet-W48 | 97.2 | 96.1 | 90.8 | 86.3 | 89.3 | 86.6 | 83.1 | 90.4 |
| HRNet-W48+ASDA | **97.3** | **96.5** | **91.7** | **87.9** | **90.8** | **88.2** | **84.2** | **91.4** |

**Compare with methods that also use parsing information.** Nie et al [19] also use parsing information and improves the 8-stacked hourglass from 90.2% to 91.0% on MPII validation set. The improvement is slightly lower than ASDA that improves the 8-stacked hourglass from 90.2% to 91.1%. In addition, [19] uses 2-stacked hourglass as Parsing Encoder to predict the parameters of an adaptive convolution, which introduces extra parameters and computation burden. Moreover, the parsing annotation and keypoints annotation of LIP are both used in the training of Parsing Encoder while our ASDA only uses the parsing annotation.

**Fig. 5.** Examples of estimated poses on the COCO test set.

## 5 Conclusions

In this work, we proposed Semantic Data Augmentation (SDA) which locally pasted segmented body parts with various semantic granularity to synthesize challenging cases. Based on the SDA, we further proposed Adversarial Semantic Data Augmentation which exploit a generative network to online adjust the augmentation parameters for each individual training image in an adversarial way. Improved results on public benchmark and comprehensive experiments have demonstrated the effectiveness of our methods. Our ASDA is general and independent on network. We hope our work can provide inspiration on how to generate tailored training samples for other tasks.

**Acknowledgement.** This work was supported by the National Natural Science Foundation of China under grant 61871435 and the Fundamental Research Funds for the Central Universities no. 2019kfyXKJC024.

## References


1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR. pp. 3686–3693 (2014)
2. Bulat, A., Tzimiropoulos, G.: Human pose estimation via convolutional part heatmap regression. In: ECCV. pp. 717–732. Springer (2016)
3. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: CVPR. pp. 7103–7112 (2018)
4. Chen, Y., Shen, C., Wei, X.S., Liu, L., Yang, J.: Adversarial posenet: A structure-aware convolutional network for human pose estimation. In: ICCV. pp. 1212–1221 (2017)
5. Chu, W., Hung, W.C., Tsai, Y.H., Cai, D., Yang, M.H.: Weakly-supervised caricature face parsing through domain adaptation. ICIP (2019)
6. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: CVPR. pp. 1831–1840 (2017)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
8. Fieraru, M., Khoreva, A., Pishchulin, L., Schiele, B.: Learning to refine human pose estimation. In: CVPR Workshops. pp. 205–214 (2018)
9. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: CVPR. pp. 932–940 (2017)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS. pp. 2672–2680 (2014)
11. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: ECCV. pp. 34–50. Springer (2016)
12. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: BMVC. vol. 2, p. 5 (2010)
13. Ke, L., Chang, M.C., Qi, H., Lyu, S.: Multi-scale structure-aware network for human pose estimation. In: ECCV. pp. 713–728 (2018)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. ICLR
15. Li, W., Wang, Z., Yin, B., Peng, Q., Du, Y., Xiao, T., Yu, G., Lu, H., Wei, Y., Sun, J.: Rethinking on multi-stage networks for human pose estimation. arXiv preprint arXiv:1901.00148 (2019)
16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV. pp. 740–755 (2014)
17. Liu, T., Ruan, T., Huang, Z., Wei, Y., Wei, S., Zhao, Y., Huang, T.: Devil in the details: Towards accurate single and multiple human parsing. arXiv preprint arXiv:1809.05996 (2018)
18. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV. pp. 483–499. Springer (2016)
19. Nie, X., Feng, J., Zuo, Y., Yan, S.: Human pose estimation with parsing induced learner. In: CVPR (2018)
20. Ning, G., Zhang, Z., He, Z.: Knowledge-guided deep fractal neural networks for human pose estimation. IEEE Transactions on Multimedia **20**(5), 1246–1259 (2018)
21. Peng, X., Tang, Z., Yang, F., Feris, R.S., Metaxas, D.: Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In: CVPR (2018)

22. Su, Z., Ye, M., Zhang, G., Dai, L., Sheng, J.: Cascade feature aggregation for human pose estimation. arXiv preprint arXiv:1902.07837 (2019)
23. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. arXiv preprint arXiv:1902.09212 (2019)
24. Tang, W., Wu, Y.: Does learning specific features for related parts help human pose estimation? In: CVPR. pp. 1107–1116 (2019)
25. Tang, W., Yu, P., Wu, Y.: Deeply learned compositional models for human pose estimation. In: ECCV. pp. 190–206 (2018)
26. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: NIPS. pp. 1799–1807 (2014)
27. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: CVPR. pp. 1653–1660 (2014)
28. Wang, X., Shrivastava, A., Gupta, A.: A-fast-rcnn: Hard positive generation via adversary for object detection. In: CVPR. pp. 2606–2615 (2017)
29. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR. pp. 4724–4732 (2016)
30. Xiao, B., Wu, H., Wei, Y.: Simple baseline for human pose estimation and tracking. In: ECCV. pp. 466–481 (2018)
31. Yang, W., Li, S., Ouyang, W., Li, H., Wang, X.: Learning feature pyramids for human pose estimation. In: ICCV. pp. 1281–1290 (2017)
32. Yu, A., Grauman, K.: Semantic jitter: Dense supervision for visual comparisons via synthetic images. In: ICCV. pp. 5570–5579 (2017)
33. Zhang, H., Ouyang, H., Liu, S., Qi, X., Shen, X., Yang, R., Jia, J.: Human pose estimation with spatial contextual information. arXiv preprint arXiv:1901.01760 (2019)