applied sciences

MDPI

Article

Self-Attention Network for Human Pose Estimation

Hailun Xia [1,2,3,*] and Tianyang Zhang [1,2,3]

[1] Beijing Key Laboratory of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications, Beijing 100876, China; zhangtianyang@bupt.edu.cn
[2] Beijing Laboratory of Advanced Information Networks, Beijing University of Posts and Telecommunications, Beijing 100876, China
[3] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
* Correspondence: xiahailun@bupt.edu.cn

Abstract: Estimating the positions of human joints from monocular single RGB images has been a challenging task in recent years. Despite great progress in human pose estimation with convolutional neural networks (CNNs), a central problem still exists: the relationships and constraints, such as symmetric relations of human structures, are not well exploited in previous CNN-based methods. Considering the effectiveness of combining local and nonlocal consistencies, we propose an end-to-end self-attention network (SAN) to alleviate this issue. In SANs, attention-driven and long-range dependency modeling are adopted between joints to compensate for local content and mine details from all feature locations. To enable an SAN for both 2D and 3D pose estimations, we also design a compatible, effective and general joint learning framework to mix up the usage of different dimension data. We evaluate the proposed network on challenging benchmark datasets. The experimental results show that our method has significantly achieved competitive results on Human3.6M, MPII and COCO datasets.

Keywords: human pose estimation; self-attention network; joint learning framework; local and nonlocal consistencies; end-to-end training

Citation: Xia, H.; Zhang, T. Self-Attention Network for Human Pose Estimation. *Appl. Sci.* **2021**, *11*, 1826. https://doi.org/10.3390/app11041826

Received: 31 January 2021
Accepted: 16 February 2021
Published: 18 February 2021

1. Introduction

Human pose estimation from monocular single images to provide informative knowledge for numerous applications, including action/activity recognition [1–3], action detection [4], human tracking [5], video gaming, surveillance, etc., is a fundamental problem in computer vision. It is a challenging problem in the presence of self-occlusions and rare poses caused by complex independent joints and high degree-of-freedom limbs, foreground occlusions caused by complex environment, etc. [6]. However, with the development of Convolution Neural Networks (CNNs) [7], significant progress has been made in in recent years. The layers of Convolution neural networks generate heat maps to represent the maximum likelihood of joints. Then, they regress these heat maps to 2D or 3D key-point locations.

Despite its good performance, we find that the convolution method is more difficult when considering anatomical relations and constraints. For example, when estimating human poses, CNNs summarize human body shapes more by texture than by geometry, and fail to capture geometrical parts of human bodies such as joint-location limits (for example, elbows between hands and shoulders) and left–right symmetry. One possible explanation for this is that CNN-based approaches rely heavily on convolution operators to model joints across the whole body shape. However, convolution operators have limited receptive fields—the long-term distance information can only be received after passing through several convolution layers. This could prevent models from learning long-range dependencies for a reason: it is difficult to summarize long-term information with a small network. As a result, parameters are too sensitive to unseen features. This will drop

high-level semantic information, which can lead to optimization not being able to achieve a better performance. Although a larger kernel size and more convolution blocks will solve this problem, they also increase computation and time efficiency.

To deal with this problem simply, effectively and efficiently, we introduce a self-attention mechanism [8] into our model which we called the Self-Attention Network (SAN). It simulates a nonlocal relationship between feature maps and combines long-term distance information into original feature maps which can significantly increase performance and efficiency of model. It is integral so the training procedure is end-to-end. SANs produce attention masks to reweight original features. These masks make the model focus more on nonlocal information. It complements convolution operators and learns contextual features and multilevel semantic information across feature maps.

This approach is general and can be used for both 2D and 3D pose estimations indistinguishably. We proposed a joint learning framework to enable the mixed usage of 2D and 3D data. Rich annotated 2D data could complement small scale 3D data. This satisfies end-to-end training and improves the generalization of model.

The main contribution of this work is three-fold:

(1) We propose a simple yet surprisingly effective self-attention approach (SAN) which exploits long-range dependency between feature maps. It can increase representation power and performance of convolution operators.
(2) We design a joint learning framework to enable usage of mixed 2D and 3D data such that the model can output both 2D and 3D poses and enhance generalization. As a by-product, our approach generates high quality 3D poses for images in the wild.
(3) In experiments, SAN advances competitive results on a 3D Human3.6M dataset [9] by a large margin and achieves 48.6 mm (Mean per joint position error (MPJPE)). On a 2D dataset, SAN achieves competitive results—91.7% (PCKh@0.5) on MPII [10] and 71.8 (AP) on COCO [11].

2. Related Work

Human pose estimation has been a widely discussed topic in the past. It is divided into 2D and 3D human pose estimations. In this section, we focus on recent learning-based methods that are most relevant to our work. We will also discuss related works on visual attention mechanisms for complementing convolution operators.

2D human pose estimation: In recent years, significant progress has been made in 2D pose estimation due to the development of deep learning and rich annotated datasets. The authors of [12] stacked bottom-up and top-down processing with intermediate supervision to improve the performance. These methods are used by many researches for 2D detection in 3D pose estimation tasks. The authors of [13] incorporated a stacked hourglass model with a multicontext attention mechanism to refine the prediction. The authors of [14] learned to focus on specific regions of different input features by combining a novel attention model. Different from these, our approach adopts a self-attention mechanism to increase the receptive field in an efficient way to learn more semantic information.

3D human pose estimation: There are two main ways to recover 3D skeleton information. The first one divides 3D pose estimation task into two su-tasks: 2D pose estimation and inference of a 3D pose from a 2D pose [15,16]. This method combines a 2D pose detector [12] and a depth regression step to estimate 3D poses. In this method, the 2D/3D poses are separated so as to generalize 3D poses in the wild images. The second one directly infers the 3D pose from RGB images [7,17,18]. In this way, the training procedure is end-to-end. We adopted this method for our approach. The authors of [17] proposed a volumetric representation for 3D poses and adopted a coarse-to-fine strategy to refine the prediction. The authors of [7] combined the benefits of regression and heat maps. We also adopted this method to make the training process differentiable and to reduce quantization error to improve network efficiency.

Mixed 2D/3D Data training: Although the end-to-end training process is concise, there is a disadvantage—the small scale of 3D in the wild annotated data limits the performance and accuracy of domain shifts. For this problem, the authors of [19] proposed a network architecture that comprises a hidden space to encode 2D/3D features. The authors of [20] proposed a weakly supervised approach to make use of large scale 2D data. The authors of [7] used soft argmax to regress 2D/3D poses directly from images. The authors of [21] proposed a method that combines 2D/3D data to compensate for the lack of 3D data. In our approach, we propose a joint learning method that separates x, y, z heat maps to mix 2D and 3D data.

Self-Attention Mechanism: When humans look at global images, they pay more attention to important areas and suppress other unnecessary information. Attention mechanisms simulate human vision and have achieved great success in computer vision—for example, scene segmentation, style transfer, image classification and action recognition. In particular, self-attention has been proposed [8] to calculate the response at a position in a sequence by attending to all positions within the same sequence. The authors of [22] proposed a cross-modal self-attention module that captures the information between linguistic and visual features. The authors of [23] assembled a self-attention mechanism into a style-agnostic framework to catch salient characteristics within images. We propose a network to extend self-attention mechanisms in human pose estimation task in feature maps to learn anatomical relationships and constraints for better recognition in nonlocal regions.

3. Model Architecture

In this section, we first describe the problem formally and give an overview of our approach. Then, we introduce the basic idea of our approach.

3.1. Overview

3D human pose estimation is a problem, where given a single RGB image or a series of RGB images $I = \{I_1, I_2, \ldots, I_i\}$, the human pose estimation process aims to localize 2D (or 3D) human body joints in Euclidean space, denoted as $Y = \{y_1, y_2, \ldots, y_k\}$, $y_k \in \mathbb{R}$ (k is the number of key-points).

As mentioned before, there are some occlusions, including the occlusion of body by objects in space and self-shielding and ambiguities including appearance diversity and lighting environment in the input. During the pose estimation stage, these weak-points will severely limit prediction ability. To solve this problem in an effective way, we adopt a new architecture, as shown in Figure 1. This is an end-to-end framework including a backbone block, self-attention network and upsampling block. Chief among them is the self-attention network which picks up efficient features from input images and generates self-attention masks to reweight the original feature maps to learn the long-range dependency between global features. A self-attention layer can capture the relatedness between feature maps and simulate long-distance multilevel associations across joints. The design of self-attention networks will be explained in Section 3.3. Backbone blocks are mainly used to extract features from image batches and upsampling blocks are used to regress feature maps to higher resolutions to refine joint locations. Backbone and upsampling block designs will be explained in Section 3.2.

Driven by the problem of lacking a 3D annotated dataset, we adopt a joint learning framework which separates x, y, z location regression in the training process—explained in Section 3.4. This method enables using mixed 2D and 3D data. It also increases the module generalization to real-world scenarios and refines the performance.

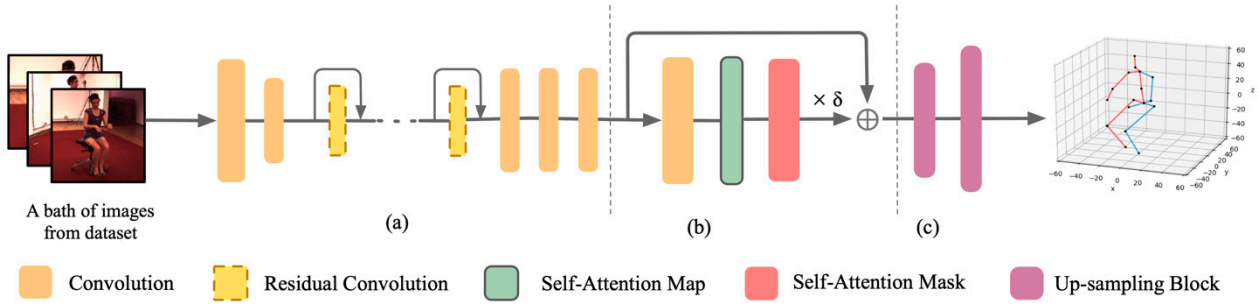

Figure 1. Framework: Illustration of the proposed approach. The basic structure contains three parts. (**a**) ResNet backbone which is used to extract features from images. (**b**) Self-Attention Network (SAN), which learns long-range dependency to compensate the lost features from original images and reweight obtained feature maps. With increasing δ, the model will depend more on nonlocal information than local content. (**c**) An upsampling block to regress the feature maps to higher resolutions to refine joint locations.

3.2. Backbone and Upsampling Block Design

In the backbone block, we adopted ResNet [24] to extract features from input images. ResNet replaces the traditional convolution + pooling layer of the deep neural network that sweeps both horizontal and vertical directions across the image. It adds a skip connection to ensure that higher layers have perform well as lower layers. Our model preserves conv1, conv2_x, conv3_x, conv4_x and conv5_x and removes the Fully Connected (FC) layer in ResNet. Because we use Resnet-50 and ResNet-152 as our backbone, the kernel size and strides are different based on network depth. In the upsampling block, we implemented deconvolution layers to regress obtained feature maps to a higher resolution. This block will refine the joint locations.

3.3. Self-Attention Network Design

When observing batches of input images including humans, we find that the relationships and constraints between joints will produce more useful information. Many human pose estimation methods use convolution neural networks (CNNs), the performance of which is limited by valid receptive field such that they are only capable of adjacent content in feature maps and cannot process long-range relations and grasp high-level semantic information. To compensate for this drawback, we propose a nonlocal approach called a Self-Attention Network (SAN). SANs not only receive efficient features in a local region, but also perceive contextual information over a wide range. The details are shown in Figure 2.

Feature maps from the previous hidden layer $X \in \mathbb{R}^{C \times B \times H \times W}$ (C is channel number, B is batch size, $H \times W$ is the pixel number) are first transformed to three feature spaces, where:

$$W_q^t = \beta_q \odot x_t, W_k^t = \beta_k \odot x_t, W_v^t = \beta_v \odot x_t \quad (1)$$

t indicates the target feature maps index. All three space vectors come from the same input. W_q^t is a query space vector. W_k^t is a key space vector. W_q^t and W_k^t are used to calculate weights which represent the similarity features between feature maps. W_v^t is a value space vector and is an output from original feature maps. Reweighting the long-term information on W_v^t enables the network to capture joint relationships easily. β_q is a weight matrix of the query space vector, which maps the input matrix of $B \times C \times W \times H$ dimensions to $B \times \frac{C}{8} \times W \times H$ dimensions. β_q needs to be transposed for the following operations. β_k is a weight matrix of the key space vector, which maps the input matrix of $B \times C \times W \times H$ dimensions to $B \times \frac{C}{8} \times W \times H$ dimensions. β_v is a weight matrix of the value space vector, which maps the input matrix of $B \times C \times W \times H$ dimensions to $B \times C \times W \times H$ dimensions. β_q, β_k and β_v are all trainable weight matrixes that transform feature maps to corresponding vector spaces and were implemented as 1×1 convolutions in our experiment.

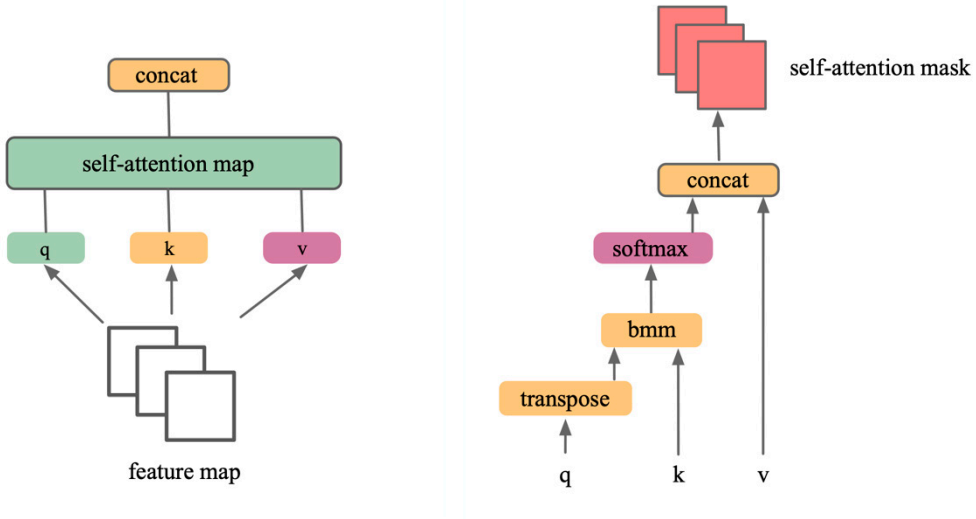

Figure 2. An illustration of a self-attention network. (**a**) SAN's inputs are three space vectors from feature maps. (**b**) Detail of self-attention map building process. The output of the SAN is a mask which will reweight the original feature maps.

W_q^t as a search vector of feature maps for one image $x \in \mathbb{R}^{C\times B\times H\times W}$ matches to the key vector W_k^t of all feature maps in this batch to calculate the positional encoding result, which is used to represent the similarity and relevance of features in the image.

$$A_t = \left[\sum_{m=1}^{t} W_q^m \odot P_m\right]^T W_k^t \tag{2}$$

A_t is the self-attention distribution and is one element of self-attention matrix $A \in \mathbb{R}^C$. m and k indicate the feature map index. A_t represents the degrees of influence of the m feature map to k feature map, whereby the model obtains any two elements dependencies of the global context. P_m is an element of a corresponding feature map, and the dimension of P_m is the same as W_q^m. In the early stage of training, the feature extraction module was not fully trained due to weight matrix and bias, and it picks up limited helpful features which will lead to a small number of A_t. So, we dropped the $1/\sqrt{d_k}$ element mentioned in [8] to reverse more adjacent information.

Then, we utilzed the softmax function on the attention mask matrix $A \in \mathbb{R}^C$ to acquire cross feature probability, which is the normalization of rows and the sum of each row after normalization is 1. The cross feature probability on *the* W_v^t space vector was reweighted to obtain W_t. W_t is a self-attention mask that captures the long-distance multilevel relationship, and considers the constraints and symmetry relationship between joints effectively.

$$W_t = \{W_1, W_2, \ldots, W_t | W_m = \frac{e^{A_i}}{\sum_{j=1}^{t} e^{A_j}}\} \otimes W_v^t \tag{3}$$

We added these self-attention masks to original feature maps with a trainable variable δ where

$$o = \delta \sum_{m=1}^{t} W_t + X \tag{4}$$

$o = \{o_1, o_2, \ldots, o_N\}$, $o \in \mathbb{R}^{C\times B\times H\times W}$ is the final output after the SAN block.

Furthermore, δ controls the ratio of the local and nonlocal features. For example, at the start of training, the network relies more on local information since it is easier. However, when time goes by, the network will assign more weight to long-term distance features to

refine the prediction. Inspired by [24], this skip connection also receives more information and mitigates the problem of a vanishing gradient.

We almost added batch normalization and ReLU at every convolution layer to speed up the training process. We used mean average loss (L1 loss) as the criterion.

3.4. Joint Learning for 2D and 3D Data

Because Equations (1)–(3) (Section 3.3) are applicable for all x, y, z coordinates in the same way, the output dimension is either 2D or 3D. So, joint learning with mixed 2D and 3D data is straightforward: separating space part x, y from depth part z. 2D data are mainly used to supervise the space part and 3D data for the depth part.

For the acquired 3D heat maps, $H_k \in \mathbb{R}^{\mathrm{W}\times\mathrm{H}\times\mathrm{D}}$ of k joints (x represents for width (W), y represents for height (H) and z represents for depth (D)). The space part H, W is always required for both 2D and 3D samples. The depth part D is only computed for 3D samples and set to 0 for 2D samples; no gradient is back-propagated from depth (D).

Taking width space x coordinate as an example, we first regressed the 3D heat map to a 1D vector:

$$I_k^x = \sum_{H_k(z)=1}^{D} \sum_{H_k(y)=1}^{H} H_k(x) \tag{5}$$

and then regressed this 1D vector into x joint location:

$$J_k^x = \sum_{H_k(x)=1}^{W} I_k^x \tag{6}$$

following this step, J_k^y and J_k^z can be inferred. In this way, the locations of x, y, z are separated so we can output 2D and 3D pose estimation results systematically.

3.5. Training and Data Processing

We used ResNet-50 and ResNet-50 as the backbone network in our experiments. The model was pretrained on an ImageNet classification dataset. δ was initialized as 0. The upsampling block for the heat map is fully convolutional. It first used deconvolution layers (4×4 kernel, stride 2) to upsample the feature map to the required resolution (72×72 for ResNet-152 and 64×64 for ResNet-50). Then, a 1×1 convolution layer was used to produce kth heat maps. Two Tesla M40 GPUs and batch size of 32 were used. The whole training contained 200 epochs. The learning rate is 0.0001 and dropped twice at the 170th epoch and 190th epoch with a decay of 0.1. An Adam optimizer was used.

In data processing, the input image was normalized to 288×384. Data augmentation included random flip, rotation ($\pm 30^\circ$), scale ($\pm 30\%$) and translation ($\pm 2\%$ of the image size) of the original image. The samples were randomly sampled and shuffled.

4. Experiment

In this section, we show our experimental results. We evaluated our model on Human3.6M 3D [9], MPII [10] and COCO [11] 2D datasets.

4.1. Dataset and Evaluation Metrics

MPII: The MPII dataset [10] is the standard benchmark for 2D human pose estimations. The images are collected from online videos covering a wide range of activities and annotated by humans for J = 16 2D joints. It contains 25,000 training images. The evaluation metric is Percentage of Correct Keypoints (PCK).

Human3.6M: The Human3.6M dataset [9] is a widely used dataset for 3D human pose estimations. This dataset contains 3.6 million RGB images captured by the MoCap System featuring 11 actors performing 15 daily activities, such as eating, sitting, walking and taking a photo, from 4 camera views. The evaluation metric is the mean per joint

position error (MPJPE), in millimeters, between the ground truth and the prediction across all cameras and joints after aligning the depth of the root joints.

COCO: The COCO dataset [11] presents imagery data with various human poses, different body scales and occlusion patterns. The training, valid and test sets contain more than 200,000 images and 250,000 in the wild person instances labels. In total, 150,000 instances are publicly available for training and valid.

4.2. Experiments on 3D Pose of Human3.6M

Following the standard protocol in [25], there are two widely used evaluation protocols with different training and testing data:

Protocol#1: Five subjects (S1, S5, S6, S7, S8) for training and two subjects (S9, S11) for testing. Mean per joint position error (MPJPE) is used for evaluation.

Protocol#2: Six subjects (S1, S5, S6, S7, S8, S9) for training and subject S11 for testing. PA mean per joint position error (PA-MPJPE) is used for evaluation.

4.2.1. Ablation Study

An ablation study was conducted using the Human3.6M test set and Protocol#1 was adopted. The self-attention network, extra 2D data, network depth and computation complexity were considered as shown in Tables 1 and 2.

Table 1. Ablative study on Human3.6M. Mean per joint position error (MPJPE) numbers are in mm.

Method	Network	Input Size	2D Data	SAN	MPJPE
a	ResNet-50	256 × 256	-	×	67.5
b	ResNet-50	256 × 256	MPII	×	62.2
c	ResNet-50	256 × 256	MPII	√	57.9
d	ResNet-50	256 × 256	MPII COCO	√	53.1
e	ResNet-50 *	288 × 384	MPII COCO	√	51.9
f	ResNet-152 *	288 × 384	MPII COCO	×	49.6
g	ResNet-152 *	288 × 384	MPII COCO	√	48.6

* indicates that flip test is used.

Table 2. Ablative study on computation complexity of models with and without SAN.

Model	Params	Flops	MPJPE
ResNet-50	34M	14.10G	62.2
ResNet-50+SAN	39M	14.43G	57.9

Self-Attention Network: Long-range dependency is important for articulated relations in human poses. Comparing to methods {*b*,*c*} and {*f*,*g*} in Table 1, performances increased by 4.3 mm MPJPE using ResNet-50 and 1 mm using ResNet-152 when considering SAN. SANs offer more long-term distance information. $\delta = 0.0$ at the beginning and increases in training. This means that long-term distance information is more important in higher level decision making processes. Considering method {*d*,*e*} in Table 1, the larger image resolution in our approach caused MPJPE to decrease by 1.2 mm. Figure 3 shows the results of the joint relationship over human joints with the change of δ. $\delta = 0.0$ means that the network has not introduced SAN, because the proximity information is easier to obtain at the beginning. Therefore, the number representing the feature gain between the joints is small. In comparison, the diagonal number is the largest, which means the current network is more dependent on neighboring features. With the increase in δ, the network introduced a more anatomical relationships and reweighted original feature maps. When $\delta = 0.8$, the larger number of means joints has strong correlation with other closed and symmetrical joints. The closed joints also have larger similarities than remote joints, such as joint 0, which has larger constraints with joint 7 than joint 15. Information transmits joints by joints such that the model will perceive more useful features. By adding this module, long-term distance information will be transformed between joints and compensate local content.

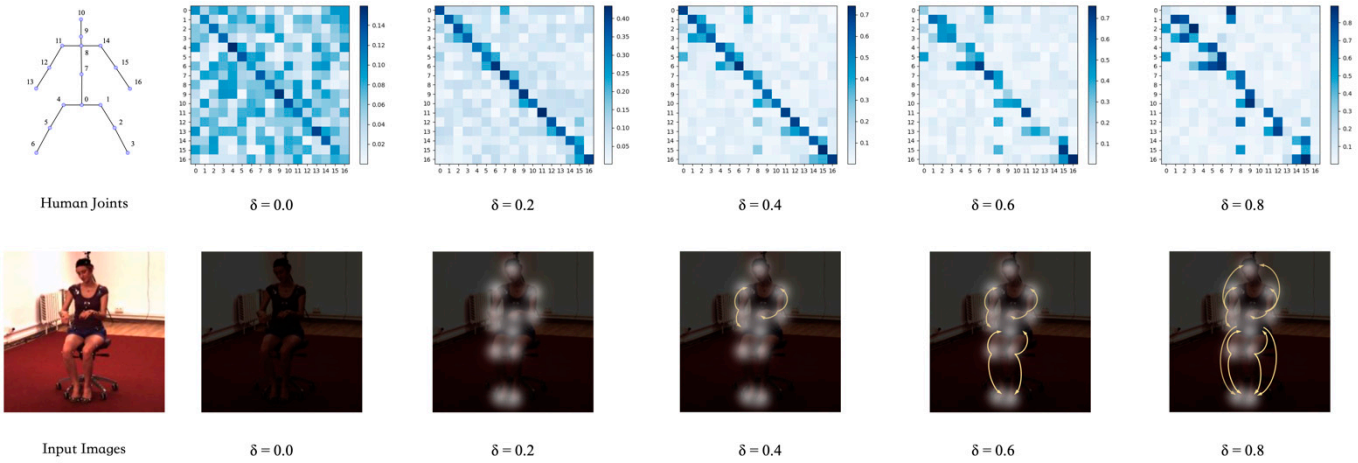

Figure 3. Results of joint relationship over human joints with the different values of δ. The first lines are confusion maps with the change of δ. Human joints show the index of joints. The confusion map represents the feature gain of the corresponding joint with other joints, and it indicates the constraints and similarities between joints. The second lines are manual connections with joints based on confusion maps.

Joint learning framework: MPII and COCO datasets provide large-scale 2D key-point in the wild data. Comparing to methods {a,b} and {c,d} in Table 1, training with both 2D and 3D data provides significant performance gain—MPJPE dropped 5.3 mm when adding MPII dataset and 4.8 mm when adopting COCO dataset. This verifies the effectiveness of joint learning framework in our training process.

Network depth: From method {e,f} in Table 1, the performance is enhanced by a deeper ResNet network. Changing network depth, MPJPE can drop by 3.3 mm from ResNet-50 to ResNet-152.

Computation complexity: Table 2 compares the model with and without SAN in terms of parameter numbers and flops(Floating-point operations per second). The parameter of the original method is 34 M and our method is 39 M. The flops of the original method are 14.10 G and our method is 14.43 G. The parameter increases 5 M and flops increases 0.33 G when adding SAN, which leads to MPJPE dropping by 4.3 mm. This verifies that the SAN model with low computation complexity will achieve a better performance.

4.2.2. Quantitative Results

The evaluation results in Table 3 show that SAN achieved good results under all protocols. Note that many leading methods have complex frameworks or learning strategies. Some of methods aim at using the wild images [19,20,32] or exploiting temporal information [28,30,33]. These methods have different research targets. Therefore, we included some of them during evaluation for completeness. There are three main findings: (1) Introducing a self-attention mechanism is effective and the proposed SAN outperforms many different type of methods in terms of results, including the end-to-end method [7,17] and two-stage method [16,19]. (2) Joint learning frameworks of 2D and 3D data are helpful [16,20]. They increase the robustness of our model in in-the-wild images. (3) Our approach showed a competitive performance on average: 48.6 mm (MPJPE) and 40.6 mm (PA-MPJPE). We improved previous methods by a large margin for the action of phones, poses, etc. The results prove the effectiveness of our approach.

Table 3. Comparison of mean per joint position error (mm) in Human3.6M between the estimated pose and the ground truth. Lower values are better, with the best in bold, and the second best underlined.

(**a**) Protocol#1: reconstruction error (MPJPE).

Protocol#1	**Direct.**	**Discuss**	**Eating**	**Greet**	**Phone**	**Pose**	**Purch.**	**Sitting**	**SittingD.**	**Smoke**	**Photo**	**Wait.**	**Walk**	**WalkD.**	**WalkT.**	**Avg.**
CoarseToFine [17]	67.4	71.9	66.7	69.1	72.0	65.0	68.3	83.7	96.5	71.7	77.0	65.8	59.1	74.9	63.2	71.9
Zhou et al. [20]	54.8	60.7	58.2	71.4	62.0	53.8	55.6	75.2	111.6	64.2	65.5	66.1	63.2	51.4	55.3	64.9
Fang et al. [26]	50.1	54.3	57.0	57.1	66.6	53.4	55.7	72.8	88.6	60.3	73.3	57.7	47.5	62.7	50.6	60.4
CompositionalHP [15]	52.8	54.8	54.2	54.3	61.8	53.6	71.7	86.7	61.5	67.2	53.1	53.4	61.6	47.1	53.4	59.1
SemanticGCN [16]	47.3	60.7	51.4	60.5	61.1	47.3	68.1	86.2	55.0	67.8	49.9	61.0	60.6	42.1	45.3	57.6
RepNet+T+M [27]	49.1	63.3	48.6	56.0	57.4	50.4	62.0	75.4	77.4	57.2	69.9	53.5	57.7	37.6	**38.1**	56.9
Propagating-LSTM [28]	43.8	51.7	48.8	53.1	52.2	52.7	**44.6**	56.9	74.3	56.7	74.9	66.4	47.5	68.4	45.5	55.8
Habibie et al. [19]	46.1	51.3	46.8	51.0	55.9	43.9	48.8	65.8	81.6	52.2	59.7	51.1	40.8	54.8	45.2	53.4
Ci et al. [18]	46.8	**38.8**	**44.7**	50.4	52.9	49.6	46.4	60.2	78.9	51.2	68.9	50.0	40.4	54.8	43.3	52.7
Li et al. [29]	43.8	48.6	49.1	49.8	57.6	45.9	48.3	62.0	73.4	54.8	61.5	50.6	43.4	56.0	45.5	52.7
Pavllo et al. [30]	47.1	50.6	49.0	51.8	53.6	49.4	47.4	59.3	67.4	52.4	61.4	49.5	39.5	55.3	42.7	51.8
Guo et al. [31]	**43.4**	50.2	48.5	**43.0**	50.6	52.4	63.8	81.1	**43.5**	61.4	**45.2**	**43.7**	55.1	**36.9**	43.5	51.8
IntegralHP [7]	47.5	47.7	49.5	50.2	51.4	43.8	46.4	58.9	65.7	49.4	55.8	47.8	38.9	49.0	43.8	49.6
Ours	43.7	44.0	47.8	48.4	**50.2**	**43.4**	46.0	**55.0**	70.9	**47.2**	52.9	44.9	**39.4**	50.6	44.1	**48.6**

(**b**) Protocol#2: reconstruction error after rigid alignment with the ground truth (PA-MPJPE), where available.

Protocol#2	**Direct.**	**Discuss**	**Eating**	**Greet**	**Phone**	**Pose**	**Purch.**	**Sitting**	**SittingD.**	**Smoke**	**Photo**	**Wait.**	**Walk**	**WalkD.**	**WalkT.**	**Avg.**
Wandt et al. [32]	53.0	58.3	59.6	66.5	72.8	56.7	69.6	78.3	95.2	66.6	71.0	58.5	63.2	57.5	49.9	65.1
Guo et al. [31]	37.8	38.9	49.7	44.7	47.3	38.9	58.5	83.6	**52.1**	62.6	**40.0**	43.2	54.2	**34.5**	39.6	48.8
CompositionalHP [15]	42.1	44.3	45.0	45.4	51.5	43.2	41.3	59.3	73.3	51.0	53.0	44.0	38.3	48.0	44.8	48.3
Fang et al. [26]	38.2	41.7	43.7	44.9	48.5	40.2	38.2	56.5	64.4	47.2	55.3	44.3	36.7	49.5	41.7	45.7
Propagating-LSTM [28]	37.4	38.9	45.6	42.6	48.5	39.9	39.2	53.0	68.5	51,5	54.6	**38.4**	33.2	55.8	37.8	45.7
Hossian et al. [33]	35.7	39.3	44.6	43.0	47.2	38.3	37.5	51.6	61.3	46.5	54.0	41.4	34.2	47.3	39.4	44.1
Li et al. [29]	35.5	39.8	**41.3**	42.3	46.0	**36.9**	37.3	51.0	60.6	44.9	48.9	40.2	33.1	44.1	**36.9**	42.6
Pavlakos [34]	**34.7**	39.8	41.8	**38.6**	**42.5**	38.0	**36.6**	50.7	56.8	42.6	47.5	39.6	**32.1**	43.9	39.5	41.8
Ours	39.7	**38.7**	43.8	42.2	43.3	38.5	42.1	**50.6**	65.9	**42.2**	46.1	39.8	33.3	44.1	38.8	**40.6**

4.2.3. Qualitative Results

Figure 4 shows the qualitative results of 3D human poses. Input images are from Human3.6M and MPII datasets. The evaluated results are accurate in both constraints and the in-the-wild environment, which shows the robustness and generalization of our model. Figure 5 shows the visualization results of failure cases. In Table 3, we can find that sitting down activity always has the worst results over other activities in many studies. The possible reason for this is that when images have serious self-occlusions, it will cause overlap between joints. The prediction accuracies of these types of activities can be improved by adding a mining difficult cases block.

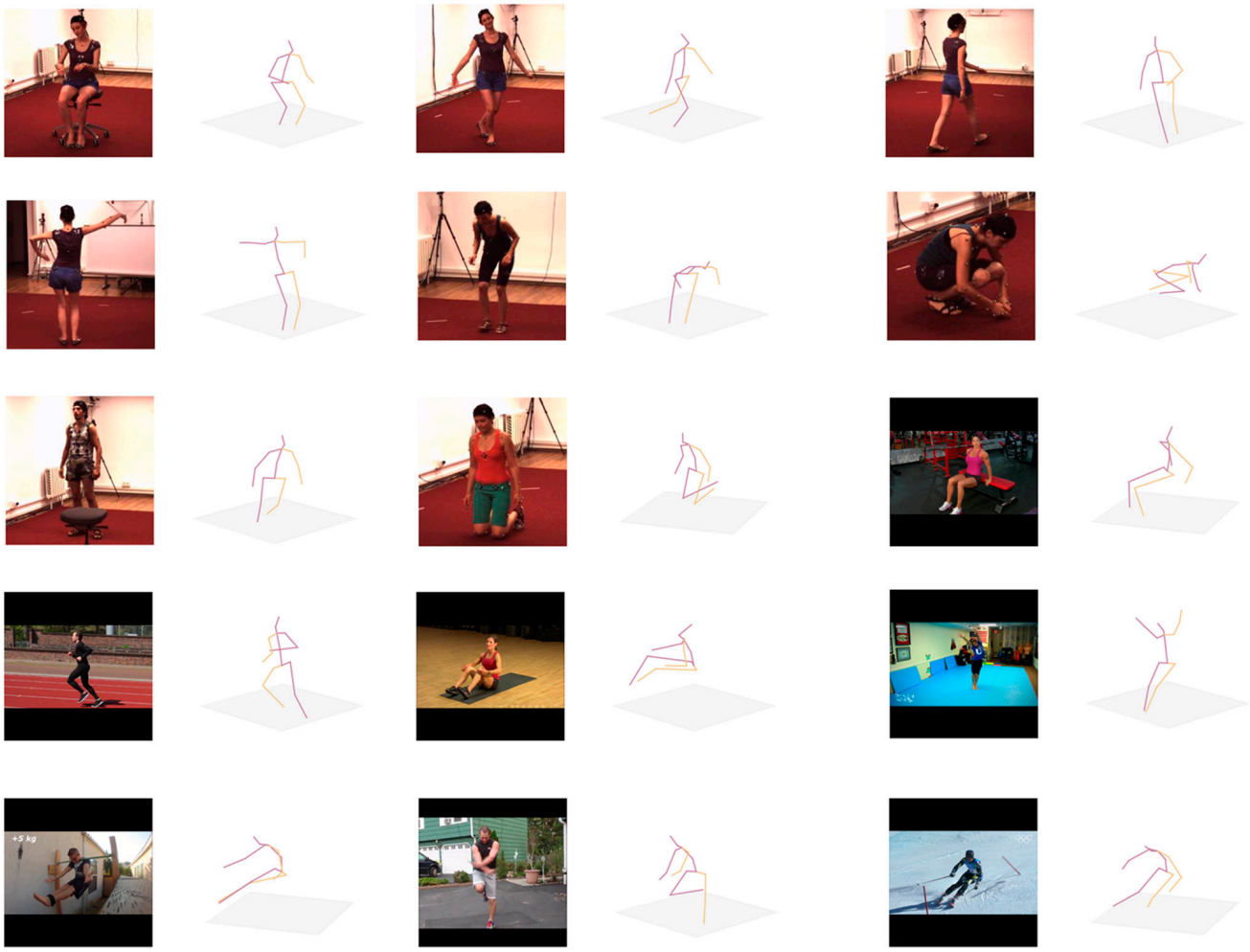

Figure 4. Qualitative results of 3D human pose on Human3.6M dataset. Predicted poses are rotated and zoomed for the consistency of perspective with original image.

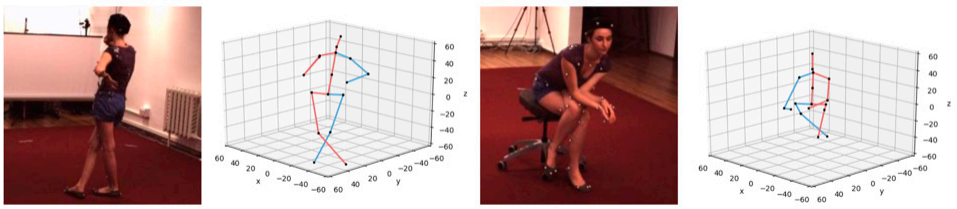

Figure 5. Failure to evaluate results of 3D human poses.

4.3. Experiment on 2D Pose of MPII and COCO

A joint learning framework enabled our model to produce high quality 2D key-point results. We carried out experiments on MPII and COCO datasets to evaluate these results. Our results were first evaluated on MPII for a validation set of about 3000 which was separate from the training, and the evaluation metric was PCK at a normalized distance of 0.5 (PCKh@0.5). Then, our results were evaluated using COCO on test-dev, and the evaluation metrics were AP, AP^{50}, AP^{75}, AP^{M} and AP^{L}.

4.3.1. Quantitative Results

Tables 4 and 5 report the comparison results of MPII and COCO, respectively. Our model achieves 91.7% (PCKh@0.5) on the MPII dataset and 71.8 (AP) on the COCO dataset, and produces competitive results and significant improvement over others. Combining the results on 3D and 2D data, we can conclude that: (1) Joint learning framework is effective. It manages 2D and 3D data in a simple way for training. (2) 2D data also increase 3D performance for rich annotation and prompt networks to produce high-quality 3D data in the wild poses.

Table 4. Comparison of PCKh@0.5 (%) on MPII. It reports the percentage of detections that fall within a normalized distance of ground truth. Higher values are better, with the best being indicated by bold font, and the second best being underlined.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Average
SimpleBaseline-152 [35]	97.0	95.9	90.3	85.0	89.2	85.3	81.3	89.6
DLCM [36]	95.6	95.9	90.7	86.5	89.9	86.6	82.5	89.8
StackedHourglass [12]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.0
PoseNFS [37]	97.9	95.6	90.7	86.5	89.8	86.0	81.5	90.2
HRNet-W32 [38]	97.1	95.9	90.3	86.4	89.1	87.1	83.3	90.3
CA+SA [39]	97.1	96.0	90.7	86.4	89.4	86.8	83.3	90.4
PRAB [39]	97.1	96.2	90.7	86.4	89.8	86.9	83.3	90.5
SkeletalHeatmap [40]	-	-	-	-	-	-	-	90.6
CU-Net [41]	97.4	96.2	91.8	87.3	90.0	87.0	83.3	90.8
DU-Net [42]	97.6	96.4	91.1	87.3	90.4	87.3	83.8	91.0
KnowledgeGuided [43]	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
MultiContext [13]	98.5	96.3	91.9	88.1	90.6	88.0	**85.0**	91.5
Ours	**98.7**	**96.6**	**92.4**	**88.2**	**91.1**	**88.9**	84.8	**91.7**

Table 5. Comparison results on COCO test-dev. Higher values are better, with the best being indicated by bold font, and the second best being underlined.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Integral-H1 [7]	66.3	88.4	74.6	62.9	72.1
Integral-I1 [7]	67.8	88.2	74.8	63.9	74.0
EfficientPose-B [44]	70.5	91.1	79.0	67.3	76.2
EfficientPose-C [44]	70.9	91.3	79.4	**67.7**	76.5
PoseNFS-MobileNet [37]	67.4	89.0	73.7	63.3	74.3
PoseNFS-ResNet [37]	70.9	90.4	77.7	66.7	**78.2**
Ours	**71.8**	**91.5**	**79.8**	66.9	78.0

4.3.2. Qualitative Results

With the help of joint learning framework, our approach outputs both 2D and 3D pose from images in the wild at the same time. We visualized example 2D prediction results in Figure 6. We can see that our method is robust in extremely difficult cases. The proposed SAN presents a better performance and can be generalized to unlimited images. This also shows that the impact of rich 2D annotated data will increase 3D performance dramatically. Our approach is helpful.

Figure 6. Qualitative results of 2D human pose on the MPII dataset.

5. Conclusions

In this paper, we propose a simple yet surprisingly effective self-attention network (SAN) for human pose estimation. SANs can not only solve the drawbacks of convolution operators, which only perceive local information and enlarge receptive fields in a computationally inefficient way, but also combine long-range dependency and multilevel information into convolution operators to enhance representation power and performance. We also introduce a joint learning framework for 2D and 3D data in the training procedure. So, our network can output both 2D and 3D poses. Experimental results show that after bringing in the SAN, the performance will be significantly improved. Our complete pipeline achieves the competitive results on both Human3.6M 3D, MPII and COCO 2D benchmarks. As a by-product, our approach generates high quality 3D poses for images in the wild.

Author Contributions: H.X., project administration, conceptualization, writing—review and editing; T.Z., investigation, methodology, writing—original draft. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (Grant No. 61976022).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: This work was funded by National Natural Science Foundation of China (Grant No. 61976022). This work is supported by the 111 project (NO.B17007).

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Li, B.; Dai, Y.; Cheng, X.; Chen, H.; Lin, Y.; He, M. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN. In Proceedings of the 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, China, 10–14 July 2017; pp. 601–604.
2. Luvizon, D.C.; Picard, D.; Tabia, H. 2D/3D Pose Estimation and Action Recognition Using Multitask Deep Learning. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5137–5146.
3. Li, B.; He, M.; Dai, Y.; Cheng, X.; Chen, Y. 3D skeleton based action recognition by video-domain translation-scale invariant mapping and multi-scale dilated CNN. *Multimed. Tools Appl.* **2018**, *77*, 22901–22921. [CrossRef]

4. Li, B.; Chen, H.; Chen, Y.; Dai, Y.; He, M. Skeleton boxes: Solving skeleton based action detection with a single deep convo-lutional neural network. In Proceedings of the 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, China, 10–14 July 2017; pp. 613–616.
5. Insafutdinov, E.; Andriluka, M.; Pishchulin, L.; Tang, S.; Levinkov, E.; Andres, B.; Schiele, B. ArtTrack: Articulated Multi-Person Tracking in the Wild. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1293–1301.
6. Chen, Y.; Tian, Y.; He, M. Monocular human pose estimation: A survey of deep learning-based methods. *Comput. Vis. Image Underst.* **2020**, *192*, 102897. [CrossRef]
7. Sun, X.; Xiao, B.; Wei, F.; Liang, S.; Wei, Y. Integral Human Pose Regression. In *Proceedings of the Constructive Side-Channel Analysis and Secure Design*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 536–553.
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Conference and Workshop on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
9. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1325–1339. [CrossRef] [PubMed]
10. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3686–3693.
11. 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 *Computer Vision—ECCV ECCV Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
12. Newell, A.; Yang, K.; Deng, J. Stacked Hourglass Networks for Human Pose Estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 483–499.
13. Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A.L.; Wang, X. Multi-context Attention for Human Pose Estimation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5669–5678.
14. Sun, G.; Ye, C.; Wang, K. Focus on What's Important: Self-Attention Model for Human Pose Estimation. *arXiv* **2018**, arXiv:1809.08371.
15. Sun, X.; Shang, J.; Liang, S.; Wei, Y. Compositional Human Pose Regression. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2621–2630.
16. Zhao, L.; Peng, X.; Tian, Y.; Kapadia, M.; Metaxas, D.N. Semantic Graph Convolutional Networks for 3D Human Pose Re-gression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3425–3435.
17. Pavlakos, G.; Zhou, X.; Derpanis, K.G.; Daniilidis, K. Coarse-to-Fine volumetric prediction for single-image 3d human pose. In Proceedings of the IEEE Proc. International Conference on Computer Vision and Pattern Recognition (CVPR), Venice, Italy, 21–26 July 2017; pp. 1263–1272.
18. Ci, H.; Wang, C.; Ma, X.; Wang, Y. Optimizing Network Structure for 3D Human Pose Estimation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 2262–2271.
19. Habibie, I.; Xu, W.; Mehta, D.; Pons-Moll, G.; Theobalt, C. In the Wild Human Pose Estimation Using Explicit 2D Features and Intermediate 3D Representations. In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 10905–10914.
20. Zhou, X.; Huang, Q.X.; Sun, X.; Xue, X.; Wei, Y. Towards 3D Human Pose Estimation in the Wild: A Weakly-Supervised Ap-proach. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 398–407.
21. Dabral, R.; Mundhada, A.; Kusupati, U.; Afaque, S.; Sharma, A.; Jain, A. Learning 3D Human Pose from Structure and Motion. In *Proceedings of the Constructive Side-Channel Analysis and Secure Design*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 679–696.
22. Ye, L.; Rochan, M.; Liu, Z.; Wang, Y. Cross-Modal Self-Attention Network for Referring Image Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 10494–10503.
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *arXiv* **2015**, arXiv:1512.03385.
24. Li, S.; Chan, A.B. 3d human pose estimation from monocular images with deep convolutional neural network. In *Proceedings Asian Conference on Computer Vision (ACCV)*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 332–347.
25. Yao, Y.; Ren, J.; Xie, X.; Liu, W.; Liu, Y.-J.; Wang, J. Attention-Aware Multi-Stroke Style Transfer. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1467–1475.
26. Fang, H.; Xu, Y.; Wang, W.; Liu, X.; Zhu, S.C. Learning Pose Grammar to Encode Human Body Configuration for 3D Human Pose Estimation. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), New Orleans, LA, USA, 2–7 February 2018; Volume 32. No. 1.
27. Chu, W.T.; Pan, Z.W. Semi-Supervised 3D Human Pose Estimation by Jointly Considering Temporal and Multiview Infor-mation. *IEEE Access* **2020**, *8*, 226974–226981. [CrossRef]

28. Lee, K.; Lee, I.; Lee, S. Propagating LSTM: 3D Pose Estimation Based on Joint Interdependency. In *Proceedings of the Constructive Side-Channel Analysis and Secure Design*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 123–141.
29. Li, C.; Lee, G.H. Generating Multiple Hypotheses for 3D Human Pose Estimation with Mixture Density Network. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9879–9887.
30. Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M. 3D Human Pose Estimation in Video with Temporal Convolutions and Semi-Supervised Training. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7745–7754.
31. Guo, Y.; Chen, Z. Absolute 3D Human Pose Estimation via Weakly-supervised Learning. In Proceedings of the 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), Macau, China, 1–4 December 2020; pp. 273–276.
32. Wandt, B.; Rosenhahn, B. RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7774–7783.
33. Hossain, M.R.I.; Little, J.J. Exploiting Temporal Information for 3D Human Pose Estimation. In *Proceedings of the Constructive Side-Channel Analysis and Secure Design*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 69–86.
34. Pavlakos, G.; Zhou, X.; Daniilidis, K. Ordinal Depth Supervision for 3D Human Pose Estimation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7307–7316.
35. Xiao, B.; Wu, H.; Wei, Y. Simple Baselines for Human Pose Estimation and Tracking. In *Proceedings of the Constructive Side-Channel Analysis and Secure Design*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 472–487.
36. Tang, W.; Yu, P.; Wu, Y. Deeply Learned Compositional Models for Human Pose Estimation. In *Proceedings of the Constructive Side-Channel Analysis and Secure Design*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 197–214.
37. Yang, S.; Yang, W.; Cui, Z. Pose Neural Fabrics Search. *arXiv* **2019**, arXiv:1909.07068.
38. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5686–5696.
39. Huo, Z.; Jin, H.; Qiao, Y.; Luo, F. Deep High-resolution Network with Double Attention Residual Blocks for Human Pose Estimation. *IEEE Access* **2020**, *8*, 1. [CrossRef]
40. Jun, J.; Lee, J.H.; Kim, C.S. Human Pose Estimation Using Skeletal Heatmaps. In Proceedings of the 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Auckland, New Zealand, 7–10 December 2020; pp. 1287–1292.
41. Tang, Z.; Peng, X.; Geng, S.; Zhu, Y.; Metaxas, D.N. CU-Net: Coupled U-Nets. In Proceedings of the British Machine Vision Conference (BMVC), Newcastle, UK, 3–6 September 2018.
42. Tang, Z.; Peng, X.; Geng, S.; Wu, L.; Zhang, S.; Metaxas, D. Quantized Densely Connected U-Nets for Efficient Landmark Localization. In *Proceedings of the Constructive Side-Channel Analysis and Secure Design*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 348–364.
43. Ning, G.; Zhang, Z.; He, Z. Knowledge-Guided Deep Fractal Neural Networks for Human Pose Estimation. *IEEE Trans. Multimedia* **2018**, *20*, 1246–1259. [CrossRef]
44. Zhang, W.; Fang, J.; Wang, X.; Liu, W. Efficientpose: Efficient human pose estimation with neural architecture search. *arXiv* **2020**, arXiv:2012.07086.