MSTPose: Learning Enriched Visual Information with Multi-Scale Transformers for Human Pose Estimation

Chengyu Wu *, Xin Wei , Shaohua Li , Ao Zhan

Posted Date: 27 June 2023

doi: 10.20944/preprints202306.1842.v1

Keywords: human pose estimation; multi-scale transformer; coordinate attention; one-dimensional coordinate vector regression

Article

MSTPose: Learning Enriched Visual Information with Multi-Scale Transformers for Human Pose Estimation

Chengyu Wu *, Xin Wei, Shaohua Li and Ao Zhan

School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China
* Correspondence: jerry916@zstu.edu.cn

Abstract: Human pose estimation is a complex detection task in which the network needs to capture the rich information contained in the images. In this paper, we propose MSTPose (Multi-Scale Transformer for human Pose estimation). Specifically, MSTPose leverages a high-resolution Convolution neural network (CNN) to extract texture information from images. For the feature maps from three different scales produced by the backbone network, each branch performs the coordinate attention operations. The feature maps are then spatially and channel-wise flattened, combined with keypoint tokens generated through random initialization, and fed into a parallel Transformer structure to learn spatial dependencies between features. As the Transformer outputs one-dimensional sequential features, the mainstream two-dimensional heatmap method is abandoned in favor of one-dimensional coordinate vector regression. The experiments show that MSTPose outperforms other CNN-based pose estimation models and demonstrates clear advantages over CNN + Transformer networks of similar types.

Keywords: human pose estimation; multi-scale transformer; coordinate attention; one-dimensional coordinate vector regression

1. Introduction

Human pose estimation is a crucial component in the field of computer vision, aiming to predict anatomical keypoints of the human body in 2D images. With the advancement of deep convolutional neural networks, the performance of pose estimation models has made significant progress. These models have gradually been applied to more complex scenarios, such as motion analysis [1–3] and human-computer interaction [4–6].

Currently, the mainstream models for pose estimation predominantly rely on CNN as encoders to extract texture features. Subsequently, the feature maps are decoded into higher-resolution sizes using methods such as heatmap-based approaches or direct keypoint regression, which has become a widely adopted paradigm in most pose estimation models. The Hourglass model [7], for instance, stacks multiple hourglass modules, where each module utilizes symmetric up-sampling and down-sampling processes combined with intermediate supervision to generate high-resolution feature maps. The HRNet [8] employs parallel branches for different resolution feature mappings while consistently maintaining the highest resolution branch. However, due to the nature of convolutional kernels, CNN exhibits local convolutional properties, restricting its ability to capture global dependencies within a limited receptive field. Although CNN excels at extracting texture features from images, it often lacks the capacity to learn spatial features effectively. As a consequence, the network fails to fully comprehend the information contained in the image. These limitations greatly constrain the potential of CNN-based models.

In recent years, Transformer [9] has achieved remarkable success in the field of natural language processing (NLP), continuously breaking records and topping various leaderboards. As a sequence-to-sequence model, Transformer exhibits strong modeling capabilities for dependencies between sequences. Furthermore, in the field of computer vision, Transformer excels at capturing

 doi:10.20944/preprints202306.1842.v1

the spatial features of images.The introduction of Vision Transformer (ViT) [10] marked the first application of Transformer to computer vision. The authors divided images into smaller patches, flattened them into sequences, and trained Transformer on these sequences. This simple yet effective approach quickly attracted the attention of many researchers. However, the high resolution of images poses computational challenges for pure Transformer methods, leading to the emergence of CNN+Transformer networks. One of the most representative models in this category is TFPose [11]. The authors employ CNN as an encoder, flatten the extracted features along the channel dimension, and feed them into Transformer. Finally, they use regression methods to predict keypoints. We believe that CNN+Transformer is a more optimal solution that leverages the strengths of both networks, striking a balance between speed and accuracy. However, mainstream CNN+Transformer models are still in their early stages, leaving room for exploration in terms of network integration and regression approaches. Consequently, the performance of both networks is not fully realized. Based on these considerations, this paper proposes a novel network architecture called MSTPose, which aims to address the limitations of existing models.

MSTPose utilizes HRNet [8] as the backbone network, where the output of the backbone network undergoes the coordinate attention operations [12]. Considering the semantic differences between different branches, the feature maps from the three branches are then fed into the MST module, which consists of a parallel Transformer structure. In the output phase, the conventional heatmap method is discarded to avoid the drawbacks of repeatedly dimensionality changes and the destruction of spatial structure when combining heatmaps with Transformer. Instead, this paper adopts the one-dimensional vector representation (VeR) validated in SimCC [13] to represent the keypoints. In summary, the MSTPose learns texture features through CNN, captures spatial features of the images through the MST module, and employs one-dimensional vector regression to preserve the position-sensitive spatial sequential mapping structure of the Transformer output. This work addresses the limitations of previous networks in insufficient image feature extraction and loss of spatial information in the Transformer output sequences. The main contributions of this paper are:

- We propose MSTPose in this study, which fully leverages the characteristics of CNN and Transformer to enable the network to learn rich visual representations, thereby significantly improving the network's modeling ability in complex scenes.
- The coordinate attention mechanism is introduced at the output location of the backbone network to obtain position-sensitive feature maps, which helps the Transformer extract spatial features from images.
- Considering the semantic differences between different branches, we propose MST module. By using a parallel structure, different-scale branches are separately fed into the Transformer for training. This allows the network to capture more complex semantic information and improve its detection ability for different instances.
- Conventional heatmap methods are discarded to overcome the drawback of repetitive dimensionality changes that disrupt the spatial structure of feature maps when combined with Transformer. Furthermore, we successfully integrate the VeR method with Transformer for the first time, resulting in improved predictive accuracy.
- In this study, we test MSTPose on the primary public benchmark datasets, COCO and MPII, and achieve better performance compared to CNN-based and CNN+Transformer networks.

2. Related Work

2.1. CNN-based Human Pose Estimation

In the field of human pose estimation, CNN-based methods have achieved tremendous success. Many early works aim to extract image features by using CNN as an encoder. DeepPose [14] firstly introduce CNN to address the problem of pose estimation, they propose a cascaded structure of deep

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2023 doi:10.20944/preprints202306.1842.v1

neural networks. In SimpleBaseline [15], the authors utilize transpose convolution in the output part of the backbone network to generate higher-resolution feature maps for better pose estimation.

Due to pose estimation is different from simple detection tasks, capturing the global dependencies between features is crucial. Varun Ramakrishna et al. [16] propose a sequential prediction algorithm that simulates the mechanism of message passing to predict the confidence of each variable (part), iteratively improving the estimation at each stage. Tompson et al. [17] utilize the structural relationships between human keypoints and incorporate the idea of Markov random fields to optimize the prediction results. Wei et al. [18] introduce the CPM (convolutional pose machines) network with VGG [19] as the backbone, employing a jointly trained multi-stage, intermediate supervision architecture to learn the dependencies between keypoints. George Papandreou et al. [20] propose a box-free system based on fully convolutional networks, learning the offsets of keypoints through a greedy decoding process and grouping keypoints into human pose instances.

However, due to the local convolutional nature of CNN, its ability to capture global dependencies is limited. Another approach is to enlarge the receptive field of feature maps, and there are various ways to achieve this, such as multi-scale fusion [21–24] and high-resolution representation [25]. Yilun Chen et al. [23] present a cascaded pyramid model to obtain multi-scale features, ultimately performing pose estimation by up-sampling to high-resolution feature maps. Bowen Cheng et al. [25] propose HigherNRNet, which utilizes transpose convolutions to obtain higher-resolution feature maps to perceive small-scale objects.

As networks become increasingly complex, there is a need for better methods to more comprehensively capture image information. Compared to previous works that solely rely on CNN, the emergence of Transformer brights new possibilities to pose estimation.

2.2. Transformer-based Human Pose Estimation

Transformer is a feed-forward network based on self-attention mechanism, which has achieved significant success in the field of NLP [26–31]. In recent years, with the introduction of Transformer into the visual domain, we have witnessed the rise of Transformer [10,32,33].

In the field of image segmentation, W. Wang et al. [34] propose a method called Attention-Guided Object Segmentation (AGOS) and Dynamic Visual Attention Prediction (DVAP) for unsupervised video object segmentation. T. Zhou et al. [35] introduce Matnet, which employs a two-stream encoder to transform surface features into mobile attention features at each stage of convolution. The bridge network is used for multi-level feature map fusion and acquisition, resulting in better segmentation results.

In the domain of object detection, N. Carion et al. [32] present the DETR model, which achieves higher detection accuracy by incorporating Transformer and employing a unique set prediction loss. To address the slow convergence speed and limited feature spatial resolution issues in [32], X. Zhu et al. [33] propose Deformable DETR, where the attention module focuses only on a small group of key sampling points around the reference, leading to improved performance.

In the field of human pose estimation, S. Yang et al. [36] introduce TransPose, using CNN as the encoder and incorporating Transformer for precise localization of human keypoints, capturing both short and long-range dependencies between keypoints. W. Mao et al. [37] propose TFPose, which builds upon [36] and employs direct regression of keypoints for pose estimation. K. Li et al. [38] develop the end-to-end PRTR model, which employs cascaded Transformer networks for direct regression of keypoints. B. Shan et al. [39] propose the MSRT network, which performs segmentation and superimposition of feature maps at different scales using the FAM module, and utilizes Transformer for keypoints decoding.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2023 doi:10.20944/preprints202306.1842.v1

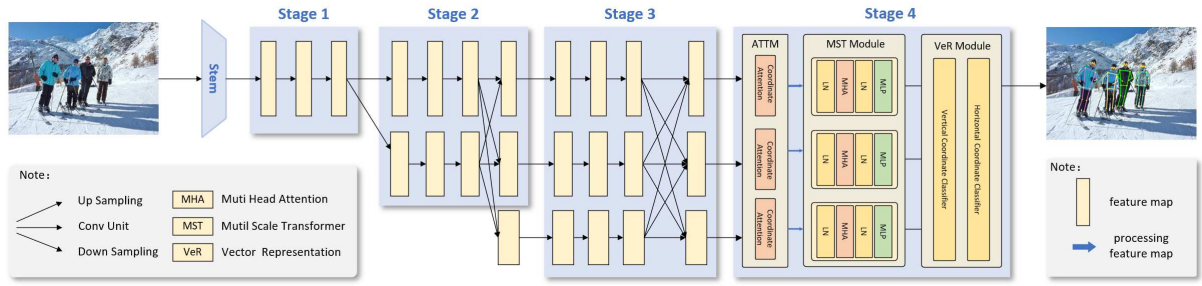

Figure 1. The structure diagram of MSTPose.

3. Proposed Method

The structure of MSTPose is illustrated in Figure 1. MSTPose employs CNN as the encoder to extract features from the images. For the feature maps of the three output branches, the coordinate attention operation [12] is applied to each branch. Subsequently, they are passed through the MST module. Finally, the keypoint coordinates are decoded using one-dimensional vector regression [13].

3.1. Backbone Network

We adopt the first three stages of HRNetW48 as the feature extractor for MSTPose, named HRNetW48-s. Assuming the input image is $X \in \mathbb{R}^{3\times H\times W}$. For the third stage of the backbone network, the outputs of each branch are $X_{1,1} \in \mathbb{R}^{C_1\times H_1\times W_1}$, $X_{1,2} \in \mathbb{R}^{C_2\times H_2\times W_2}$, and $X_{1,3} \in \mathbb{R}^{C_3\times H_3\times W_3}$, where $H_1, W_1 = H/4, W/4$, $H_2, W_2 = H/8, W/8$, and $H_3, W_3 = H/16, W/16$. These feature maps currently contain rich texture information.

3.2. ATTM

Subsequently, the outputs $X_{1,i} \in \mathbb{R}^{C_i\times H_i\times W_i}, i \in (1,2,3)$ of the backbone network are fed into the Attention module (ATTM). ATTM consists of three parallel coordinate attention mechanisms, as illustrated in Figure 2.

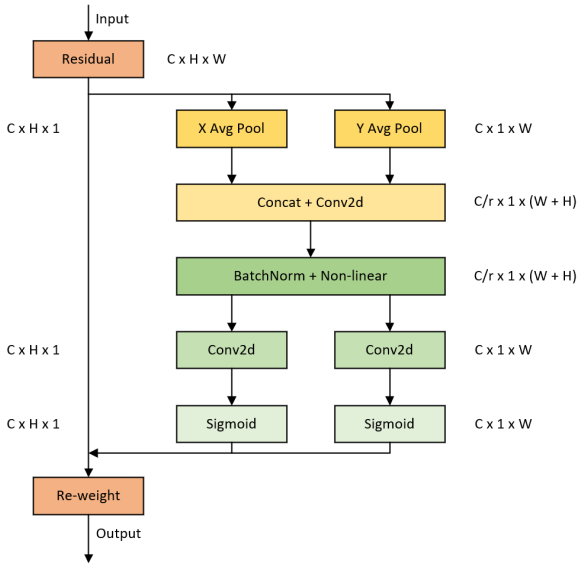

Figure 2. The module schematic diagram of the coordinate attention mechanism.

Specifically, the input feature maps are first subjected to two one-dimensional pooling operations, where the feature maps are aggregated separately along the vertical and horizontal directions, resulting in two distinct direction-aware feature maps. These two feature maps, which embed direction-specific

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2023 doi:10.20944/preprints202306.1842.v1

information, are then encoded into two attention maps. Each attention map captures long-range dependencies along the spatial direction of the input feature maps. The attention maps are then applied to the input feature maps through multiplication, resulting in direction-aware and position-sensitive feature maps. These feature maps aid in more accurate object localization and perception of objects of interest by the network. The outputs of ATTM are denoted as $X_{2,i} \in \mathbb{R}^{C_i \times H_i \times W_i}, i \in (1,2,3)$.

3.3. MST Module

Next, the outputs $X_{2,i}$ of the ATTM module are fed into the MST module, whose structure are illustrated in Figure 3. As the Transformer is a sequential network, it is necessary to map $X_{2,i}$ into one-dimensional sequences. Since visual features encompass both texture and spatial information, in order to enable the network to fully capture visual information, this paper respectively expands $X_{2,i}$ along the spatial and channel dimensions to generate channel tokens and spatial tokens. At this point, their shapes are $X_{c,i} \in \mathbb{R}^{C_i \times L_c}$ and $X_{s,i} \in \mathbb{R}^{C_i \times L_s}$, where $i \in (1,2,3)$. Meanwhile, the network initializes and generates learnable keypoint vectors as keypoints tokens, with the shape of $X_k \in \mathbb{R}^{M \times L}$, where M represents the number of keypoints labeled for each human instance, and L denotes the length of each sequence, set to 192 in this paper.

For channel tokens and spatial tokens, the network controls the length of each token to be $L = 192$ through the linear mapping. Consequently, the newly generated tokens have shapes of $X_{c,i} \in \mathbb{R}^{C_i \times L}$, $X_{s,i} \in \mathbb{R}^{C_i \times L}$, where $i \in (1,2,3)$. The advantage of this approach is that it significantly reduces the computational complexity of Transformer $O(C \times L^2)$, where C denotes the quantity of tokens and L denotes the length of tokens, while preserving fine-grained information for each sequence. Since the self-attention mechanism in the Transformer lacks positional awareness, position encoding is then applied to the channel tokens and spatial tokens. After encoding, the three types of tokens are concatenated and jointly fed into the Transformer for feature learning. The concatenated input sequences are denoted as $X_T \in \mathbb{R}^{C_T \times L}$, where $C_T = M + C_i + C_i, i \in (1,2,3)$.

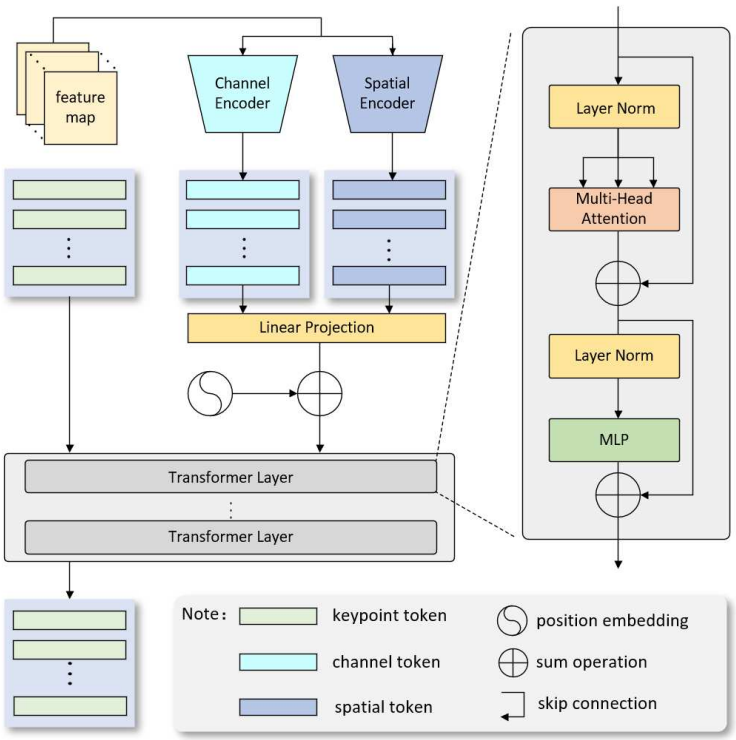

Figure 3. Schematic diagram of an individual branch in the MST module.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2023 doi:10.20944/preprints202306.1842.v1

In Transformer, the first step is to perform linear projections on the input sequences to generate Q (query), K (key), and V (value). The specific formula is shown as follows:

$$Q = X_T \times W_Q, K = X_T \times W_K, V = X_T \times W_V \tag{1}$$

where W_Q, W_K, and W_V represent the corresponding weight matrices. Then, the spatial dependencies of the features are captured using the multi-head self-attention mechanism, with the following formula:

$$MHSA(Q, K, V) = softmax(\frac{Q \times X^T}{\sqrt{L}}) \tag{2}$$

where L represents the dimension of the keys, each query needs to be paired with all the keys. Subsequently, softmax is employed to compute attention scores, with each score determining the attention level of the current query token.

3.4. VeR Module

As shown in Figure 4, in the VeR module, we directly process the one-dimensional sequences output by the Transformer to preserve its fine-grained information to the maximum extent. In this paper, we first fuse the one-dimensional sequences output from three branches to generate $X_V \in \mathbb{R}^{M \times L}$, where the number of sequences is M. Then, they are separately fed into the X and Y vector classifiers, generating $X_x \in \mathbb{R}^{M \times (H \cdot K)}$ and $X_y \in \mathbb{R}^{M \times (W \cdot K)}$. Due to the existence of the scale factor k, the lengths of the generated X_x and X_y sequences will be larger than the original image's width and height, thereby achieving sub-pixel level localization. Subsequently, X_x and X_y are decoded to generate predicted coordinates, as shown in the following formulas:

$$O_x^i = argmax[x_0, x_1, \cdots, x_{W \cdot (k-1)}] \in \mathbb{R}^{W \cdot k}, x_i = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(i - x')^2}{2\sigma^2}) \tag{3}$$

$$O_y^j = argmax[y_0, y_1, \cdots, y_{W \cdot (k-1)}] \in \mathbb{R}^{W \cdot k}, y_j = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(j - y')^2}{2\sigma^2}) \tag{4}$$

where σ represents the standard deviation, (x', y') refers to the pixel point on the (X_i, Y_j) vector, (x_i, y_j) denotes the supervised signal generated through the one-dimensional Gaussian distribution, and (O_x^i, O_y^j) represents the predicted coordinates of the keypoint.

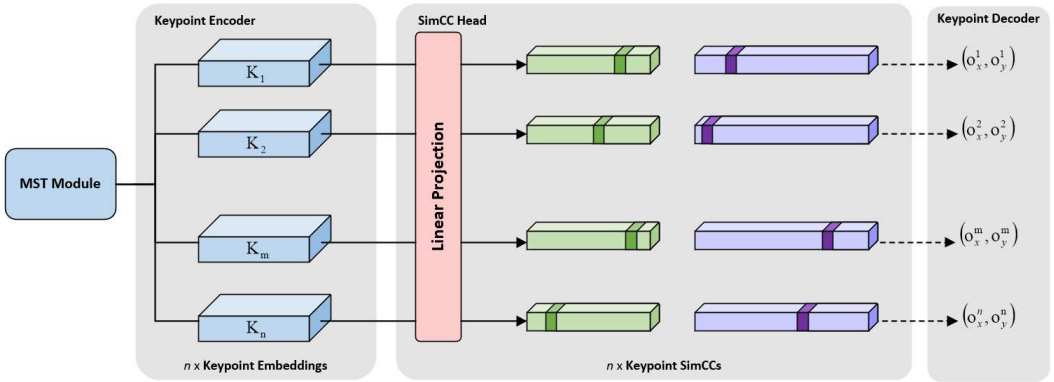

Figure 4. Schematic diagram of the VeR module.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2023 doi:10.20944/preprints202306.1842.v1

4. Experiments

4.1. Experimental Details

4.1.1. Datasets and Evaluation Indicators

Our experiments utilize the widely adopted benchmark datasets, COCO [40] and MPII [41]. In order to verify the performance of MSTPose, we extensively train and validate the model on these two datasets. The following sections provide the detailed introduction to each dataset.

COCO dataset: COCO is a large-scale and versatile dataset widely used in the field of computer vision, which is proposed by Microsoft. It consists of 200k images and 250k annotated human instances, with each human instance labeled with 17 keypoints. In this paper, we train our model on the COCO train2017 set and perform validation and ablation experiments on the COCO Validation set, which contains 5k images. We also test our model on the COCO test-dev set, consisting of 20k images, and compare its performance with state-of-the-art models. The evaluation metrics used in the COCO dataset are average precision (AP) and average recall (AR), which are derived based on the Object Keypoint Similarity (OKS). The formula for OKS is as follows:

$$OKS = \frac{\sum_i exp^{(-\frac{d_i^2}{2s^2k_i^2})}\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{5}$$

where d_i^2 represents the square of the Euclidean distance between the ground truth and predicted values, i denotes the i-th keypoint, s^2 represents the square of the area occupied by the human instance, k_i represents a constant that controls the attenuation for each keypoint, v_i indicates the visibility of the keypoint, and δ represents a logical function. The formula for AP is shown below:

$$AP_t = \frac{\sum_p \delta(OKS > t)}{\sum_p 1} \tag{6}$$

when t is set to 0.5 and 0.75, they are denoted as AP^{50} and AP^{75}, respectively. When $32^2 < s < 96^2$ and $s > 96^2$, they are denoted as AP^M and AP^L, and the same applies to AR.

MPII dataset: The MPII dataset is one of the most commonly used benchmarks in the field of human pose estimation. It consists of a total of 25k images and 40k human instances, with each instance annotated with 16 keypoints. The evaluation metric used in the MPII dataset is PCK (Percentage of Correct Keypoints), which is calculated using the following formula:

$$PCK_\sigma^p(d_0) = \frac{1}{|\tau|}\sum_\tau \delta(||x_p^f - y_p^f||_2 < \sigma) \tag{7}$$

where d_0 represents a human detector, and σ denotes a threshold that indicates the degree of match between the ground truth and predicted values.

4.1.2. Implementation Details

This article follows a top-down paradigm. Firstly, single-person images are detected from multiple-person images using a human body detector [42], followed by single-person keypoint detection. During the training process, the total number of epochs is set to 210, with an initial learning rate of 1e-3, which is reduced to 1e-4 at the 90th epoch and further reduced to 1e-5 at the 120th

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2023 doi:10.20944/preprints202306.1842.v1

epoch. The experiments are conducted on a system equipped with four NVIDIA GeForce RTX 3090 24G GPUs.

4.2. Experimental Results

4.2.1. Quantitative Experimental Results

The test results of MSTPose on the COCO Validation dataset are shown in Table 1. It can be observed that MSTPose achieves the best results in the major metrics. In comparison to TFPose[37], MSTPose exhibits decrease of 39.7% in GFLOPs, yet it surprisingly improves the AP by 4.8%. Furthermore, when compared to PRTR[38], MSTPose achieves 3.9% increase in AP while utilizing only 38.6% of PRTR[38]'s GFLOPs. Additionally, compared to the MSRT[39] network of the same type, AP also increased by 5%, showing a significant improvement. The excellent performance of MSTPose on the COCO Validation dataset confirms its feasibility.

Table 1. The testing results of MSTPose on COCO Validation dataset.The best result is bolded.

Method	Backbone	GFLOPs	Input Size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Simple Baseline[15]	ResNet50	8.9	256×192	70.4	88.6	78.3	67.1	77.2
Simple Baseline[15]	ResNet101	12.4	256×192	71.4	89.3	79.3	68.1	78.1
Simple Baseline[15]	ResNet152	15.7	256×192	72.0	89.3	79.8	68.7	78.9
TFPose[37]	ResNet50	20.4	384×288	72.4	-	-	-	-
PRTR[38]	ResNet101	33.4	512×348	72.0	89.3	79.4	67.3	79.7
PRTR[38]	HRNetW32	21.6	384×288	73.1	89.3	79.4	67.3	79.8
PRTR[38]	HRNetW32	37.8	512×348	73.3	89.2	79.9	69.0	80.9
MSRT[39]	ResNet101	-	512×348	72.2	89.1	79.2	68.1	79.4
MSTPose	HRNetW48	14.6	256×192	**77.2**	**92.9**	**84.1**	**73.9**	**81.7**

The test results of MSTPose on COCO test-dev are shown in Table 2, the MSTPose outperforms pure CNN-based networks. Furthermore, compared to a human pose estimation model that utilizes Transformer, MSTPose demonstrates remarkable competitiveness. In comparison to TFPose[37], despite 5.8% decrease in GFLOPs, MSTPose achieves 2.5% improvement in AP. When compared to PRTR[38], MSTPose achieves 2.6% increase in AP while 23.2% decrease in GFLOPs. These results highlight the significant advantages of MSTPose in terms of both speed and accuracy, whether compared to pure CNN-based networks or Transformer-based human pose estimation networks.

Table 2. The testing results of MSTPose on COCO test-dev dataset, where T is Transformer.The best result is bolded.

Method	Backbone	GFLOPs	Input Size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
DeepPose[14]	ResNet101	7.7	256×192	57.4	86.5	64.2	55.0	62.8	-
DeepPose[14]	ResNet152	11.3	256×192	59.3	87.6	66.7	56.8	64.9	-
CenterNet[42]	Hourglass	-	-	63.0	86.8	69.6	58.9	70.4	-
DirectPose[43]	ResNet50	-	-	62.2	86.4	68.2	56.7	69.8	-
PointSetNet[44]	HRNetW48	-	-	68.7	89.9	76.3	64.8	75.3	-
Integral Pose[45]	ResNet101	11.0	256×256	67.8	88.2	74.8	63.9	74.0	-
TFPose[37]	ResNet50+T	20.4	384×288	72.2	90.9	80.1	69.1	78.8	74.1
PRTR[38]	HRNetW48+T	-	-	64.9	87.0	71.7	60.2	72.5	78.8
PRTR[38]	HRNetW48+T	21.6	384×288	71.7	90.6	79.6	67.6	78.4	79.4
PRTR[38]	HRNetW48+T	37.8	512×384	72.1	90.4	79.6	68.1	79.0	-
MSTPose	HRNetW48+T	14.6	256×192	**74.7**	**91.9**	**81.7**	**71.4**	**80.1**	**79.8**

The test results of MSTPose on the MPII dataset are shown in Table 3. It can be observed that except for Hip, MSTPose achieves the best performance in all other metrics. Specifically, the Mean@0.5 score

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2023 doi:10.20944/preprints202306.1842.v1

reaches 90.2%, which is a notable improvement of 2 percentage points compared to PRTR-R101[38]. This represents a significant enhancement in the MPII dataset.

Table 3. The testing results of MSTPose on MPII dataset.The best result is bolded.

Method	Backbone	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Mean
Simple Baseline[15]	ResNet50	96.4	95.3	89.0	83.2	88.4	84.0	79.6	88.5
Simple Baseline[15]	ResNet101	96.9	95.9	89.5	84.4	88.4	84.5	80.7	89.1
Simple Baseline[15]	ResNet152	97.0	95.9	90.0	85.0	89.2	85.3	81.3	89.6
HRNet[8]	HRNetW32	96.9	**96.0**	90.6	85.8	88.7	86.6	82.6	90.1
MSRT[39]	ResNet101	97.0	94.9	89.0	84.0	**89.6**	85.7	80.3	89.1
PRTR-R101[38]	ResNet101	96.3	95.0	88.3	82.4	88.1	83.6	77.4	87.9
PRTR-R152[38]	ResNet152	96.4	94.9	88.4	82.6	88.6	84.1	78.4	88.2
MSTPose	HRNetW48	**97.1**	**96.0**	**90.8**	**86.8**	89.5	**86.8**	**82.8**	**90.2**

4.2.2. Qualitative Experimental Results

From Figure 5, it can be observed that MSTPose is capable of accurately predicting human keypoints in various occlusion scenarios, including self-occlusion and mutual occlusion. Furthermore, in densely populated scenes with indistinct human features, MSTPose performs effectively by extracting rich information embedded in the images. This enables accurate identification of each individual instance and their skeletal structure. In the top-left corner of the image, the person sitting on the far left experiences severe mutual occlusion, but MSTPose is capable of reconstructing the overall target using local features as much as possible.

Figure 5. Visualization of the testing results of MSTPose on COCO Validation dataset.

4.3. Ablation Experiments

In order to enable the network to learn rich visual information, this paper proposes the MSTPose, which is achieved through the ATTM, the MST module, and the VeR module. To enhance the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2023 doi:10.20944/preprints202306.1842.v1

persuasiveness of the network, extensive ablation experiments are conducted in this study to validate the effectiveness of each module.

4.3.1. Ablation Experiment of ATTM

The mainstream pose estimation networks based on HRNet utilize the output of the highest resolution branch as the final output of the entire backbone network, disregarding the other branches. However, MSTPose considers each branch by applying coordinate attention to each of them and then feeding them into the parallel Transformer module. To verify the effect of coordinate attention in each branch, ablation experiments are conducted as shown in Table 4. Due to the presence of multiple network branches, there are a total of eight possible combinations. Among them, we select five representative combinations for the ablation experiments. While controlling other variables, all methods employ HRNetW48-s as the backbone network and utilize the MST module and VeR module.

Table 4. The ablation experiment of coordinate attention mechanism, where CA is the coordinate attention.

Method	Branch1	Branch2	Branch3	*AP*
CA				76.7
CA	✓			77.0
CA		✓		77.0
CA			✓	76.9
CA	✓	✓	✓	77.2

It can be observed that the best performance is achieved when the coordinate attention mechanism is applied to the highest resolution branch, contributing an improvement of 0.3% *AP* for the network. Branch2 contributes 0.3% *AP*, and Branch3 contributes 0.2% *AP* to the network. When all three branches adopt the coordinate attention mechanism, the entire network achieves 0.5% *AP* improvement. Therefore, it can be concluded that the coordinate attention mechanism works better for branches with higher resolutions, which explains why previous works often focused on using only the highest resolution branch. Although branches with lower resolutions make a smaller contribution, the overall performance improvement is significant when the coordinate attention mechanism is applied to the entire network, thus affirming the effectiveness of coordinate attention for each branch.

4.3.2. Ablation Experiment of MST Module

We follow the approach outlined in 4.3.1 to conduct ablation experiments on the parallel structure of Transformer modules. The results are shown in Table 5. It can be observed that when only Branch1 is trained with Transformer, it contributes 0.6% *AP* to the network, Branch2 contributes 0.4% *AP*, and Branch3 contributes 0.4% *AP*. This further confirms previous findings that favored using the highest resolution as the network's output. Although the branch with the highest resolution contributes significantly to the network, when combined with the low resolution branches, the network achieves 0.9% *AP* improvement, demonstrating a remarkable enhancement. Hence this validates the effectiveness of the parallel branch Transformer.

Table 5. The ablation experiment of MST module.

Method	Branch1	Branch2	Branch3	*AP*
Transformer				76.3
Transformer	✓			76.9
Transformer		✓		76.7
Transformer			✓	76.7
Transformer	✓	✓	✓	77.2

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2023 doi:10.20944/preprints202306.1842.v1

4.3.3. Ablation Experiment of VeR Module

In this study, the MSTPose adopts the one-dimensional vector regression approach to predict keypoints, abandoning the conventional heatmap method. To demonstrate the superiority of the one-dimensional vector regression method and affirm the suitability of the VeR approach for human pose estimation based on Transformer networks, the comparative experiment is conducted between the two methods, as shown in Table 6.

Table 6. The ablation experiment of VeR module.

Method	Backbone	VeR	Heatmap	*AP*
method1	HRNetW48-s	✓		77.2
method2	HRNetW48-s		✓	75.1

When using the heatmap method, the *AP* is 75.1%. However, when employing the VeR method, the *AP* significantly increases to 77.2%, indicating a noticeable improvement of 2.1 percentage points. This clear enhancement supports the author's earlier claim that utilizing a one-dimensional vector representation for keypoints, based on the one-dimensional sequences output from the Transformer, is a better choice compared to the heatmap method.

4.3.4. Ablation Experiment of MSTPose

The preceding sections involve ablation experiments conducted on each individual module. Subsequently, the focus shifts to the overall performance, and the results of ablation experiments on the entire MSTPose are presented in Table 7. Comparing method1 with method2, we observe that without using the MST module, ATTM contributes *AP* of 0.4%. Comparing method3 with method4, when the MST module is utilized, ATTM contributes *AP* of 0.5%. Comparing method1 with method3, in the absence of ATTM, the MST module contributes *AP* of 0.8%. Comparing method2 with method4, when ATTM is employed, the MST module contributes *AP* of 0.9%.

Table 7. The ablation experiment of MSTPose.

Method	ATTM	MST Module	VeR	*AP*
method1			✓	75.9
method2	✓		✓	76.3
method3		✓	✓	76.7
method4	✓	✓	✓	77.2

From these findings, it is evident that ATTM, the MST module, and the VeR module significantly contribute to the network. The coordination among different components facilitates the enhancement of the network's ability to extract complex features, thereby further improving its overall performance.

5. Conclusion

In this paper, we propose a human pose estimation network based on a multi-scale parallel structure. We apply coordinate attention operations to three branches of the backbone network's output. Subsequently, these branches are fed into Transformer modules. Finally, we discard the conventional heatmap-based approach and instead adopt coordinate vector regression to predict the final keypoints. Remarkably, our method achieves satisfactory results. We conduct extensive tests on mainstream datasets, validating the outstanding performance of MSTPose. Additionally, we perform numerous ablation experiments to verify the effectiveness of each module.

Author Contributions: Conceptualization, Chengyu Wu, Xin Wei and Shaohua Li; Data curation, Xin Wei; Formal analysis, Xin Wei; Funding acquisition, Chengyu Wu; Investigation, Ao Zhan; Methodology, Xin Wei; Project administration, Chengyu Wu; Software, Shaohua Li; Supervision, Ao Zhan; Validation, Chengyu Wu;

Visualization, Shaohua Li; Writing – original draft, Shaohua Li; Writing – review & editing, Chengyu Wu and Ao Zhan.

Funding: This research was funded by the First Batch of "Pioneer" and "Leading Goose" R&D Programs of Zhejiang Province in 2023 under grant 2023C01041.

Data Availability Statement: Publicly archived datasets used in the study are listed below. COCO: http://cocodataset.org (accessed on 15 November 2022); MPII: http://human-pose.mpi-inf.mpg.de (accessed on 10 March 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Meng, Z.; Zhang, M.; Guo, C.; Fan, Q.; Zhang, H.; Gao, N.; Zhang, Z. Recent Progress in Sensing and Computing Techniques for Human Activity Recognition and Motion Analysis. *Electronics* **2020**, *9, 1357*.
2. Agostinelli, T.; Generosi, A.; Ceccacci, S.; Khamaisi, R.K.; Peruzzini, M.; Mengoni, M. Preliminary Validation of a Low-Cost Motion Analysis System Based on RGB Cameras to Support the Evaluation of Postural Risk Assessment. *Appl. Sci.* **2021**, *11, 10645*.
3. Maskeliūnas, R.; Damaševičius, R.; Blažauskas, T.; Canbulut, C.; Adomavičienė, A.; Griškevičius, J. BiomacVR: A Virtual Reality-Based System for Precise Human Posture and Motion Analysis in Rehabilitation Exercises Using Depth Sensors. *Electronics* **2023**, *339*.
4. Liu, Hai, et al. ARHPE: Asymmetric relation-aware representation learning for head pose estimation in industrial human–computer interaction. *IEEE Transactions on Industrial Informatics* **2022**, *18.10, 7107-7117*.
5. Liu, Hai, et al. Precise head pose estimation on HPD5A database for attention recognition based on convolutional neural network in human-computer interaction. *Infrared Physics & Technology* **2021**, *116, 103740*.
6. Wang, K., Zhao, R. and Ji, Q., 2018, May. Human computer interaction with head pose, eye gaze and body gestures. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG, 2018, pp. 789-789.
7. Newell, Alejandro, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14, pp. 483-499.
8. Sun, K., Xiao, B., Liu, D. and Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Long Beach, USA, June 16-20, 2019, pp. 5693-5703.
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. Attention is all you need. Advances in neural information processing systems, Long Beach, USA, December 4th, 2017, 30.
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2020, arXiv:2010.11929.
11. Mao, W., Ge, Y., Shen, C., Tian, Z., Wang, X. and Wang, Z. Tfpose: Direct human pose estimation with transformers. arXiv 2021, arXiv:2103.15320.
12. Hou, Qibin, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Online, June 19-25, 2021, pp. 13713-13722.
13. Li, Y., Yang, S., Liu, P., Zhang, S., Wang, Y., Wang, Z., Yang, W. and Xia, S.T., 2022, November. SimCC: A Simple Coordinate Classification Perspective for Human Pose Estimation. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI, pp. 89-106.
14. Toshev, A., & Szegedy, C. Deeppose: Human pose estimation via deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, Columbus, OH, USA, June 23-28, 2014, pp. 1653-1660.
15. Xiao, B., Wu, H. and Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the European conference on computer vision, Munich, Germany, September 8-14, 2018, pp. 466-481.

16. Ramakrishna, V., Munoz, D., Hebert, M., Andrew Bagnell, J. and Sheikh, Y. Pose machines: Articulated pose estimation via inference machines. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13, pp. 33-47.
17. Tompson, J.J., Jain, A., LeCun, Y. and Bregler, C. Joint training of a convolutional network and a graphical model for human pose estimation. Advances in neural information processing systems, Long Beach, USA, December 4th, 2017, 27.
18. Wei, S.E., Ramakrishna, V., Kanade, T. and Sheikh, Y. Convolutional pose machines. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, USA, June 26-July 1, 2016, pp. 4724-4732.
19. Simonyan, K. Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556.
20. Papandreou, G., Zhu, T., Chen, L.C., Gidaris, S., Tompson, J. and Murphy, K. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In Proceedings of the European conference on computer vision. Munich, Germany, September 8-14, 2018, pp. 269-286.
21. Pfister, T., Charles, J. and Zisserman, A. Flowing convnets for human pose estimation in videos. In Proceedings of the IEEE international conference on computer vision. Santiago, Chile, December 7-13, 2015, pp. 1913-1921.
22. Yang, W., Li, S., Ouyang, W., Li, H. and Wang, X. Learning feature pyramids for human pose estimation. In proceedings of the IEEE international conference on computer vision. Venice, Italy, October 22-29, 2017, pp. 1281-1290.
23. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G. and Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, USA, June 18-22, 2018, pp. 7103-7112.
24. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L. and Wang, X. Multi-context attention for human pose estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition. Hawaii, USA, July 21-26, 2017, pp. 1831-1840.
25. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S. and Zhang, L. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle, USA, June 14-19, 2020, pp. 5386-5395.
26. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv 2018, arXiv:1810.04805.
27. Liu, Y. et al. Roberta: A robustly optimized bert pretraining approach. arXiv 2019, arXiv:1907.11692.
28. Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. Improving language understanding by generative pre-training. 2018.
29. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv 2019, arXiv:1910.13461.
30. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 2020, 21(1), pp. 5485-5551.
31. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V. and Mihaylov, T. Opt: Open pre-trained transformer language models. arXiv 2022, arXiv:2205.01068.
32. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S., 2020. End-to-end object detection with transformers. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, pp. 213-229.
33. Zhu, X., Su, W., Lu, L., Li, B., Wang, X. and Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. arXiv 2020, arXiv:2010.04159.
34. Wang, W., Song, H., Zhao, S., Shen, J., Zhao, S., Hoi, S.C. and Ling, H., 2019. Learning unsupervised video object segmentation through visual attention. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach, USA, June 16-20, 2019, pp. 3064-3074.
35. Zhou, T., Li, J., Wang, S., Tao, R. and Shen, J. Matnet: Motion-attentive transition network for zero-shot video object segmentation. IEEE Transactions on Image Processing, 2020, 29, pp. 8326-8338.

36. Yang, S., Quan, Z., Nie, M. and Yang, W., 2021. Transpose: Keypoint localization via transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, October 10-17, 2021, pp. 11802-11812.
37. Mao, W., Ge, Y., Shen, C., Tian, Z., Wang, X. and Wang, Z. Tfpose: Direct human pose estimation with transformers. arXiv 2021, arXiv:2103.15320.
38. Li, K., Wang, S., Zhang, X., Xu, Y., Xu, W. and Tu, Z., 2021. Pose recognition with cascade transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online, June 19-25, 2021, pp. 1944-1953.
39. Shan, B., Shi, Q. and Yang, F. MSRT: multi-scale representation transformer for regression-based human pose estimation. Pattern Analysis and Applications, 2023, 26(2), pp. 591-603.
40. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll´ar, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. 2014, pp. 740–755.
41. Andriluka, M., Pishchulin, L., Gehler, P. and Schiele, B., 2014. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. Columbus, USA, June 23-28, 2014, pp. 3686-3693.
42. Zhou, X., Wang, D., & Krähenbühl, P. Objects as points. arXiv 2019, arXiv:1904.07850.
43. Tian, Z., Chen, H., & Shen, C. Directpose: Direct end-to-end multi-person pose estimation. arXiv 2019, arXiv:1911.07451.
44. Wei, F., Sun, X., Li, H., Wang, J. and Lin, S., 2020. Point-set anchors for object detection, instance segmentation and pose estimation. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16, pp. 527-544.
45. Sun, X., Xiao, B., Wei, F., Liang, S. and Wei, Y. Integral human pose regression. In Proceedings of the European conference on computer vision. Munich, Germany, September 8-14, 2018, pp. 529-545.