electronics

Article

A Multi-Channel Parallel Keypoint Fusion Framework for Human Pose Estimation

Xilong Wang [1], Nianfeng Shi [2,*], Guoqiang Wang [2], Jie Shao [1] and Shuaibo Zhao [3]

[1] College of Electronics and Information Engineering, Shanghai University of Electric Power, Shanghai 201306, China; wangxilong@mail.shiep.edu.cn (X.W.); shaojie@shiep.edu.cn (J.S.)
[2] School of Computer and Information Engineering, Luoyang Institute of Science and Technology, Luoyang 471023, China; wgq@lit.edu.cn
[3] College of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China; 220320040446@stu.haust.edu.cn
* Correspondence: shinf@lit.edu.cn

Abstract: Although modeling self-attention can significantly reduce computational complexity, human pose estimation performance is still affected by occlusion and background noise, and undifferentiated feature fusion leads to significant information loss. To address these issues, we propose a novel human pose estimation framework called DatPose (deformable convolution and attention for human pose estimation), which combines deformable convolution and self-attention to relieve these issues. Considering that the keypoints of the human body are mostly distributed at the edge of the human body, we adopt the deformable convolution strategy to obtain the low-level feature information of the image. Our proposed method leverages visual cues to capture detailed keypoint information, which we embed into the Transformer encoder to learn the keypoint constraints. More importantly, we designed a multi-channel two-way parallel module with self-attention and convolution fusion to enhance the weight of the keypoints in visual cues. In order to strengthen the implicit relationship of fusion, we attempt to generate keypoint tokens to the visual cues of the fusion module and transformers, respectively. Our experimental results on the COCO and MPII datasets show that performing the keypoint fusion module improves keypoint information. Extensive experiments and visual analysis demonstrate the robustness of our model in complex scenes and our framework outperforms popular lightweight networks in human pose estimation.

Keywords: human pose estimation; deformable convolution; self-attention; keypoint fusion; lightweight networks

Citation: Wang, X.; Shi, N.; Wang, G.; Shao, J.; Zhao, S. A Multi-Channel Parallel Keypoint Fusion Framework for Human Pose Estimation. *Electronics* **2023**, *12*, 4019. https://doi.org/10.3390/electronics12194019

Academic Editor: KC Santosh

Received: 21 August 2023
Revised: 18 September 2023
Accepted: 21 September 2023
Published: 24 September 2023

1. Introduction

Estimating the 2D coordinates of human keypoints from images is a fundamental research topic in the field of computer vision. This has a broad application prospect, including human activity recognition [1], action quality evaluation [2], and autonomous driving [3]. It requires consideration of both the position information and the constraint relationships between the keypoints.

Recent studies have achieved remarkable success in human pose estimation by spatially locating keypoints alone [4–6]. However, these methods rely on scale information to achieve high-resolution data, which requires significant computational resources. Additionally, feature extraction based on a fixed size of convolution and pooling kernels cannot effectively capture the constraint relationships between keypoints. These constraints represent the interdependencies and geometric relationships between different body parts. For example, the position of the elbow determines the position of the wrist, and the alignment of the neck affects the tilt of the head. The human pose is a complex system, with fixed relationships between its components. These relationships can be based on anatomical geometric constraints or on dynamic constraints related to movement and actions. These

constraint relationships influence the position, angles, and relative positions of various body parts. However, traditional fixed-size convolution and pooling operations are not suitable for capturing the constraint relationships between keypoints. This is because fixed-size operations cannot adaptively handle variations in poses, angles, and relative positions. They treat each keypoint as an independent entity and overlook the interdependencies and geometric constraints between keypoints. This can lead to suboptimal accuracy in pose estimation, as the relationships between keypoints are not fully utilized. Therefore, developing a robust model that can effectively recognize and establish relationships among keypoints is crucial for accurate human pose estimation. To achieve this, researchers must focus on improving the model's ability to emphasize essential keypoint information.

Researchers have introduced the transformer model, originally used in natural language processing (NLP) [7], to advocate research in this direction. Enforcing the vision transformer for visual cues constraints is an innovative and effective method for pose estimation [8–10]. The transformer model utilizes a self-attention mechanism in its encoder and decoder modules, enabling it to calculate the response by considering all location features in the feature map weighted. This inherent global modeling capability has led to significant advancements in various pose estimation tasks, as evidenced by the numerous transformer-based models. Yang et al. [8] introduced a method that leverages image tokens to capture visual cues, akin to the way word2vec captures similarity between words and characters in a vector space. However, although their embedded attention mechanism is capable of computing global attention, it overlooks the crucial constraint relationship between keypoints and visual cues. Therefore, Li et al. [9] proposed a new method called TokenPose to solve this problem. Specifically, TokenPose introduces the utilization of tokens to represent individual keypoints. This approach facilitates the acquisition of both visual cues and constraint relations through interactions with visual and other keypoint tokens. While the constraint strategy effectively addresses the limitations of fusing visual cues and keypoint information, it does introduce some background noise. Additionally, the keypoint tokens are treated together with visual cues, without strengthening keypoint information.

In this work, we propose a novel convolution and self-attention parallel multi-channel keypoint fusion method, which aims to emphasize keypoint features. Some works, such as Transpose and HRformer [8,11], are based on convolution neural network (CNN) as a back-bone, utilizing early layers to capture low-level visual information and deeper layers for richer feature expression. However, in DatPose, the situation is quite distinct. Our primary objective in designing the Deformable Convolution is to selectively capture edge keypoint features specific to the human body in an adaptive way. In the first stage, rather than simply extracting visual cues, we extract two streams of features in parallel using convolution and attention mechanisms to strengthen the key-point information. Finally, we divide the feature map into patches and keypoints as tokens, which are fed into the Transformer encoder to learn the constraint relationship between visual cues and keypoints, thus improving the network's performance.

The main contributions of this paper can be summarized as follows:

(1) We introduce a deformable convolution that can selectively adjust the target of a human body image, reducing information redundancy by filtering out irrelevant information and placing it in the appropriate location.

(2) We propose a keypoint fusion module that combines convolution and self-attention to enhance keypoint information and minimize background noise.

(3) Experimental results on COCO demonstrate that our proposed method, DatPose, efficiently incorporates information from visual cues and keypoint information at multiple levels, achieving state-of-the-art performance on 2D metrics.

The present research is organized as follows: Section 2 provides a comprehensive overview of existing literature in the field, Section 3 elaborates on the architecture of DatPose, Section 4 presents the experimental validation and in-depth analysis, and finally, the paper concludes with pertinent findings and conclusions.

2. Related Work

The subsequent passage presents a concise overview of pertinent literature on vision transformers, 2D pose estimation, and convolution-enhanced attention.

2.1. Vision Transformer

The transformer architecture was initially introduced in the natural language processing domain to overcome the issue of long-distance dependencies and has resulted in significant advancements in classification, segmentation, detection, and virtual reality. Recently, the Vision Transformer [12] has been adapted to computer vision by splitting images into patches and processing them as tokens, akin to NLP inputs. Liu et al. [13] introduced a hierarchical architecture that incorporates the fusion of image patches in deeper layers. This design enables the model to effectively process images with diverse dimensions. It also introduced a shift window mechanism that computes self-attention in non-overlapping windows locally. Various transformer-based models have undergone enhancements through widely used model compression techniques such as Deit [14], which employed knowledge distillation methods to acquire inductive biases inherent in CNNs. Nevertheless, these approaches primarily concentrate on particular classification tokens and are not directly applicable to pose estimation tasks. In contrast, Rao et al. [15] employed a dynamic token sparsification framework to progressively and dynamically remove redundant tokens.

2.2. 2D Human Pose Estimation

Two-dimensional pose estimation has witnessed significant progress in recent years, with CNN architectures being the typical solution for human pose estimation [4]. Unlike 3D human pose estimation [16], these architectures use a multi-scale approach to capture keypoint information by changing the resolution through the use of hourglass structures. However, this approach may not fully exploit information from various scales. In this regard, Sun et al. [17] and Wu et al. [18] achieved high accuracy by parallel convolutional extraction of features from different resolutions while maintaining a high resolution. Nonetheless, the method is computationally expensive and does not consider the constraints between keypoint information. Xu et al. [10] leveraged transformer-based methods to deal with these spatial constraints. As an extension, Yang et al. [8] combined convolutional and transformer-based methods to further improve performance. Nonetheless, such methods may be vulnerable once keypoints are partially obscured, as their constraints may be insufficiently strong. To mitigate this issue, Li et al. [9] proposed a separate keypoint extraction mechanism, later integrated with visual information to enhance the inter-keypoint constraints. However, this approach treats visual cues and keypoint information equally, without considering the greater importance of keypoint information in visual cues. In response, we propose a novel method that combines deformable convolution and transformer-based approaches to better capture the significance of keypoints in visual cues.

2.3. Convolution Enhanced Attention

In computer vision tasks, especially in vision transformers, the self-attention network's inductive bias is weak. To address this issue, several methods have introduced convolution operations to enhance the capability of inductive bias. Wu et al. [19] employed convolution in the tokenization process and integrated stride convolution to reduce the computation complexity of self-attention. ViT [12] with convolutional stem achieved better performance by adding convolutions at the early stage. Dong et al. [20] introduced positional coding based on convolution and showcased advancements in downstream tasks. Additionally, Peng et al. [21] merged a transformer with a separate CNN model to incorporate both features. However, existing approaches often integrate features from cascade hierarchies, whereas our method strives to eliminate such cascade dependencies and process features in a parallel way, aligning better with the transformer's objective of reducing computational amount. Furthermore, in contrast to the conventional approach of augmenting the high-

level features generated by deep convolutional neural networks with fine-grained low-level features, our proposed fusion attention module specifically targets keypoint feature information. This emphasis on keypoint feature integration distinguishes our method from others. We integrate the keypoint information into the convolutional stream, allowing for joint learning and increasing the weight of keypoint information relative to visual cues.

3. Materials and Methods

Figure 1 depicts the overall architecture of our proposed DatPose, which employs convolution and self-attention blocks to extract keypoints at the human body edges. Initially, in order to mitigate the intricacy involved in subsequent feature extraction and acquire a feature map F with dimensions $H \times W \times C$, where H, W, and C represent height, width, and channel, respectively, we introduce image I as the input to the stem CNN. To enhance the keypoint information, we introduce a fusion block to increase the ratio of keypoints to visual cues, which is referred to as the fusion of convolution and self-attention. Specifically, we divide the feature map into two streams: the convolution stream and the attention stream. The convolution layer multiplies the keypoints to acquire local keypoint information, while the self-attention layer learns the global visual cues and the constraints between key-points. Finally, the two streams are combined into a feature map. We divide the fused feature map and input it to the Transformer encoder to learn global dependencies. This multi-stage approach reinforces the keypoint information.

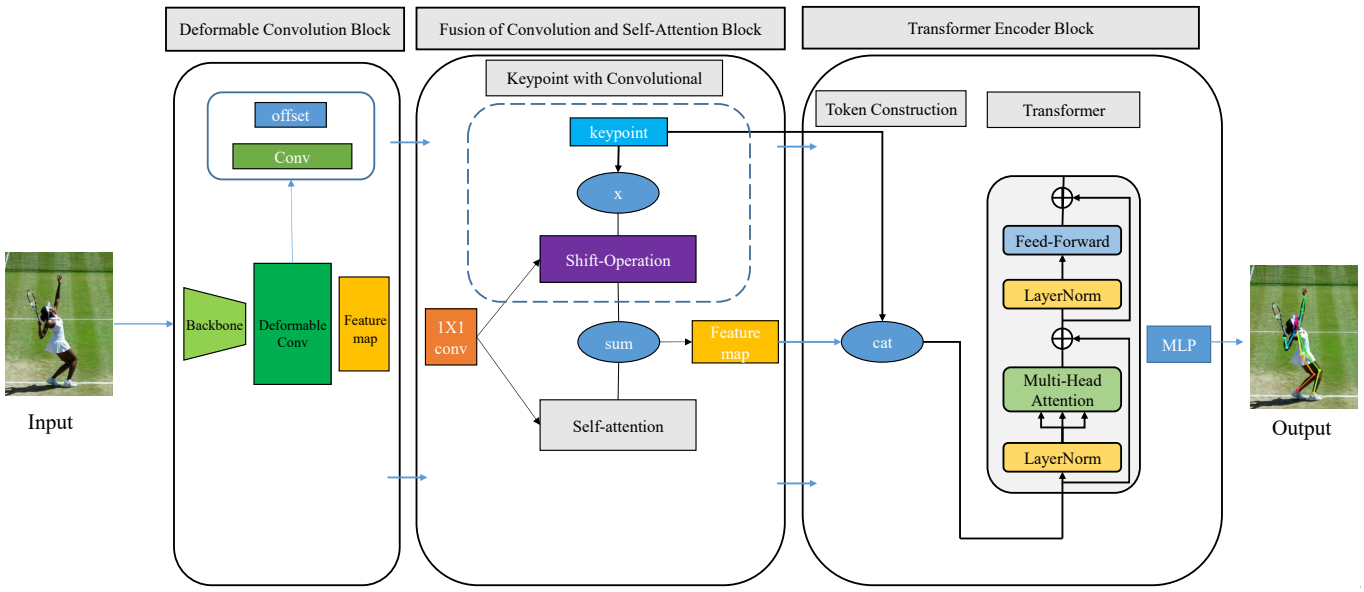

Figure 1. An overview of our model. The model contains three modules: the deformable convolution block aims to capture keypoints of human body edge and the fusion of convolution and self-attention block supports the keypoint information and visual cues weight distribution. Furthermore, the Transformer encoder conducts token construction and constraint relationship learning.

3.1. Deformable Convolution

Deformable convolution is well-known for feature extraction and offset learning [22,23]. The 2D convolution can be formulated as:

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p) \quad (1)$$

where $w(p_n)$ is the weight matrix applied to the feature map $x(p_0 + p_n + \Delta p)$ and $p_n + \Delta p$ represents the offset locations. The regular grid R is augmented with offsets $\{\Delta p_n \mid n = 1, \ldots, N\}$, where p_n enumerates the location in R and $N = |R|$.

To ensure the accurate pixel position, the Formula (1) can be written in the following form

$$x(p) = \sum G(q, p) \cdot x(q) \tag{2}$$

where p denotes an arbitrary fractional location ($p_0 + p_n + \Delta p$); the sum symbol $\sum$ denotes the sum of all the terms of the source pixel position q. Each term is composed of the weight function G (q, p) multiplied by the value x(q) of the corresponding source pixel position q. By summing all terms, the value x(p) of the target pixel position p can be obtained. It is used to ensure that each source pixel position in the formula takes into account the contribution of the target pixel position. q enumerates all integral spatial locations in the feature map x, and G (·,) is the bilinear interpolation kernel. To ensure the accurate pixel position, bilinear interpolation is performed to achieve the position offset.

$$G(q,p) = g(q_x, p_x)g\left(q_y, p_y\right) \tag{3}$$

where g (a, b) = max(0, 1 − | a − b |).

By utilizing this deformable convolution operation, the feature map can dynamically adapt to the specific shape of the target, which is beneficial to capture the keypoints of the human body edge.

3.2. Fusion of Convolution and Self-Attention

The essence of pose estimation is effectively aggregating relevant keypoint information while filtering out irrelevant visual information. Treating keypoint information and visual cues equally by using linear layers is not a prudent approach. We propose a fusion module that enhances keypoint information in the presence of visual cues. This module consists of two streams: the keypoint with convolution stream and the attention stream, which is the core of pose estimation.

3.2.1. Keypoint with Convolution

To overcome the interference of irrelevant visual information and enhance the keypoint information, we propose a fusion block. The fusion block consists of two essential components: keypoint elementary and visual cues. We regard the convolution operation as a summation of shifted feature maps and achieve it by using three 1×1 convolutions. These convolutions refer to the use of 1×1-sized kernel filters in the convolution operation. These 1×1 convolutions can be employed to change the number of channels in a feature map, providing a way to transform the representation of information. The formula for the operation is:

$$g_{ij} = \sum_{p,q} K_{p,q} f_{i+p-\lfloor k/2 \rfloor, j+q-\lfloor k/2 \rfloor} \tag{4}$$

Consider a standard convolution with a kernel $K \in \mathcal{R}^{C_{out} \times C_{in} \times k \times k}$, where k represents the kernel size and C_{in} and C_{out} denote the input and output channel sizes, respectively. Where $K_{p,q} \in \mathcal{R}^{C_{out} \times C_{in}}$ and the indices p and q range from 0 to $k - 1$, representing the kernel weights associated with the kernel position (p, q). For convenience, we can rewrite as the summation of the feature maps from different kernel positions:

$$g_{ij} = \sum_{p,q} g_{ij}^{(p,q)} \tag{5}$$

In the above formulation, to simplify the formulation further, we introduce the Shift operation as $\widetilde{f} \triangleq \text{Shift}(f, \Delta x, \Delta y)$, which represents shifting the feature map f by Δx units in the horizontal direction and Δy units in the vertical direction as:

$$\widetilde{f} \triangleq \text{Shift}(f, \Delta x, \Delta y) \tag{6}$$

where Δx, Δy correspond to the horizontal and vertical displacements. Then, the formulation can be rewritten as:

$$g_{ij} = \text{Shift}\left(K_{p,q}f_{ij}, p - \lfloor k/2 \rfloor, q - \lfloor k/2 \rfloor\right) \tag{7}$$

Based on the formulation, the convolution kernel $K_{p,q}f_{ij}$ is applied to the input of the position $(p - \lfloor k/2 \rfloor, q - \lfloor k/2 \rfloor)$ by applying the Shift operation to obtain the output g_{ij}. In order to enhance the representation and importance of keypoint information in the convolution flow, the keypoint information X_k is introduced, which contains k keypoints, and the keypoint information is integrated into the convolution flow by multiplying X_K with the convolution kernel $K_{p,q}f_{ij}$ element by element by using the '$*$' operation. The keypoint with convolution can be formulated as:

$$g_{ij} = \text{Shift}\left(X_k * K_{p,q}f_{ij}, p - \lfloor k/2 \rfloor, q - \lfloor k/2 \rfloor\right) \tag{8}$$

where k represents the N keypoints, which add out channels. Specifically, each keypoint information X_k is multiplied by the elements at the corresponding position of the input feature map$K_{p,q}f_{ij}$. In this way, the elements of the keypoint information corresponding to the position will be amplified or weakened, thereby enhancing the weight of the keypoint. The '$*$' operation makes the keypoints obtain higher weights throughout the convolution process. This makes the information of key points more prominent than visual cues, as shown in Figure 2. According the operation, the keypoint can obtain more weight compared to the visual cues.

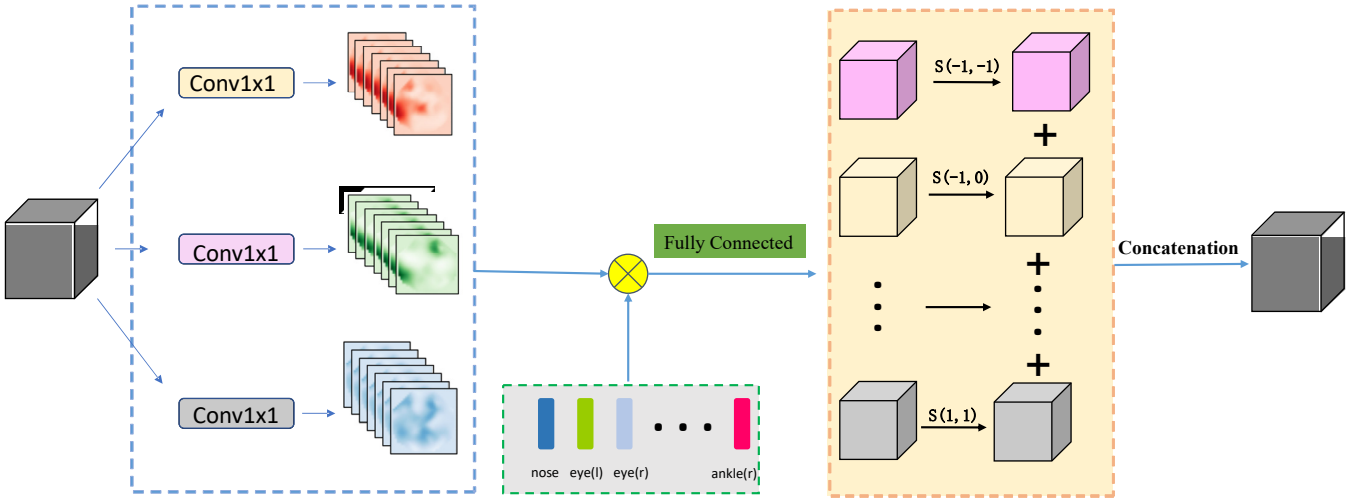

Figure 2. An illustration of the proposed shift operation. The feature map is projected with three 1×1 convolutions and the intermediate features are multiplied by the keypoints. s (x, y) corresponds to the shift operation defined in Formula (7). $\otimes$ denotes the elementwise multiplication operation.

3.2.2. Fusion of Self-Attention Mechanism

The input of self-attention is same as the keypoint with convolution, separated by the three 1×1 convolutions. As shown in Figure 3, three given inputs: Query Q, Key K, and Value V of the same dimension Q, K, V, give the output which is computed as a weighted sum

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{Softmax}\left(\frac{\text{QK}^T}{\sqrt{d_k}}\right)\text{V} \tag{9}$$

where the parameters inside the activation function Softmax(.) reflect the similarity of Q and K. To avoid the resulting small gradients affecting the training, d_k is the dimension of tokens, the d_k is usually used to scale the size of the QK^T. The self-attention mechanism can reflect the contribution of different image positions through gradients [24–26].

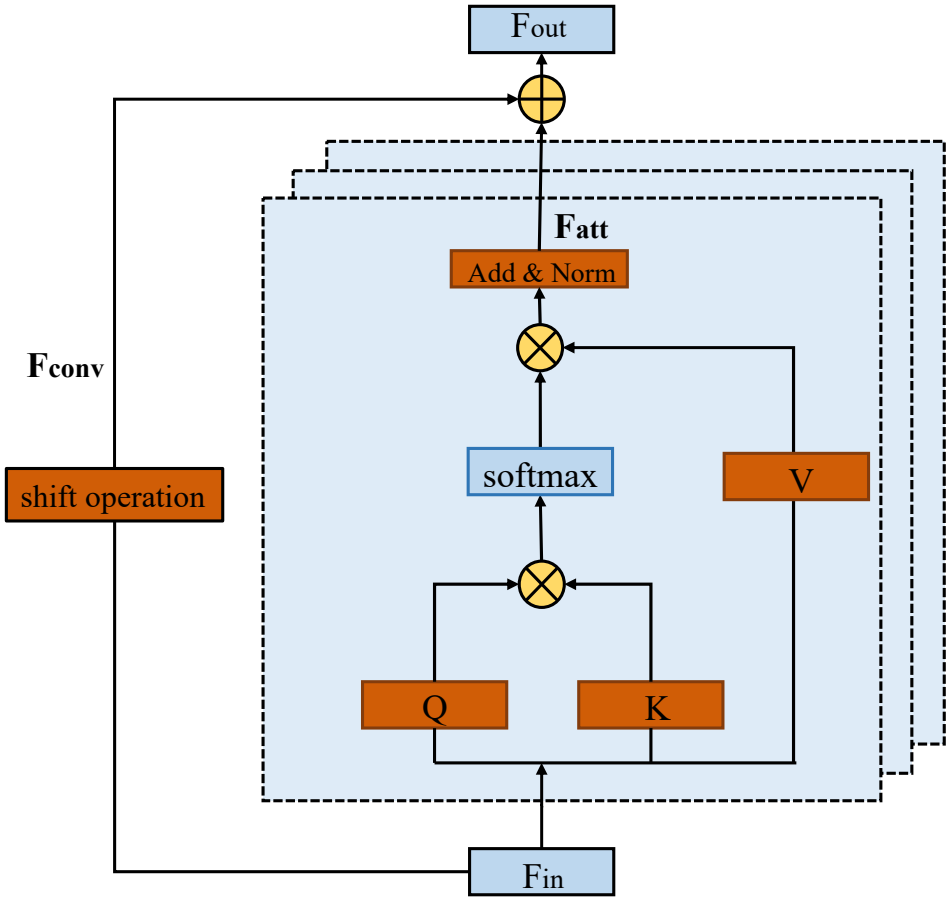

Figure 3. Illustration of the proposed fusion module. Given feature map fin, the shift operation represents that the Q, K, and V are embedded by the three kernel size is (1 × 1). ⊕ denotes the element-wise addition. ⊗ denotes the matrix multiplication.

In our study, we employ the ACmix [27] approach, where the two paths are added and fused to achieve our final result. Furthermore, we also utilize the learned scalar to regulate the intensity of the convolution with the aim of enhancing keypoint information.

$$F_{\text{out}} = \alpha F_{\text{att}} + \beta F_{\text{conv}} \tag{10}$$

3.3. Transformer Module

In order to accurately predict the location information of human keypoints, we propose a joint approach that integrates visual information with keypoint information, allowing for mutual interaction to improve the performance of human target detection, even under low resolution. We use the Transformer model, known for its ability to capture dependencies between elements, to facilitate the robust detection and tracking of keypoints. Specifically, we segment the feature map into several patches, which are then encoded using the Transformer model. Finally, the multi-layer perceptron (MLP) model is employed to predict the keypoints. This joint approach offers a promising solution for enhancing the effectiveness of keypoint detection in human targets.

3.3.1. Construction of Token

After constructing feature maps by combining convolution and self-attention layers, the feature maps are split into visual and keypoint tokens, as shown in Figure 4. The visual token, denoted by x, captures constraints among the visual tokens, while the keypoint token is designed to learn the constraints between keypoints, which helps to address

low-resolution and occluded keypoints. These tokens are concatenated and fed into the Transformer Encoder to learn the dependencies between tokens.

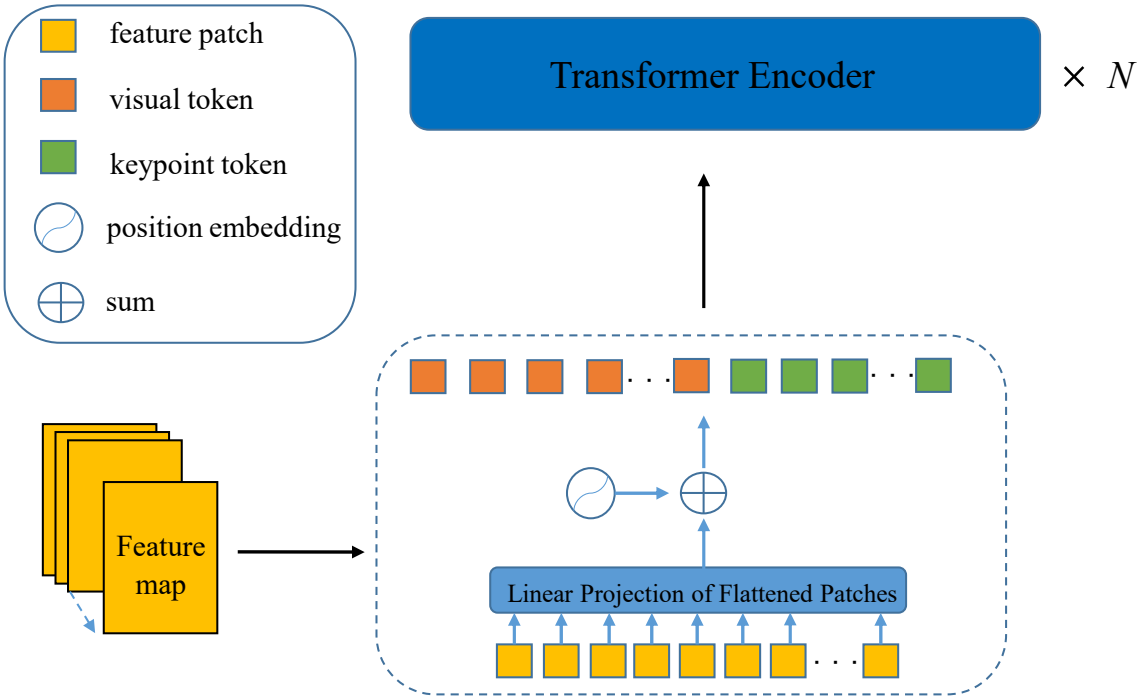

Figure 4. Construction of token. The feature map x is divided into N patches, which are then transformed into a 1D vector through the linear projection of the flattened patches layer. The vector that is created in one dimension is utilized as a visual token, followed by position encoding that incorporates a sine strategy. The result is then combined with keypoints through concatenation.

3.3.2. Transformer Encoder

Given a 1D token as the input of the Transformer, which consists of N Transformer modules, each module contains a multi-head self-attention module and a multi-class prediction module. Layer Norm [28] is applied to each module. The core formula of the Transformer is as follows:

$$\mathrm{SA}(\mathrm{T}^{\mathrm{l}-1}) = \mathrm{softmax}(\frac{\mathrm{T}^{\mathrm{l}-1}\mathrm{W}_{\mathrm{Q}}(\mathrm{T}^{\mathrm{l}-1}\mathrm{W}_{\mathrm{K}})^{\mathrm{T}}}{\sqrt{\mathrm{d}_{\mathrm{h}}}})(\mathrm{T}^{\mathrm{l}-1}\mathrm{W}_{\mathrm{V}}) \tag{11}$$

where W_{K}, W_{V}, and W_{Q} are parameters that belong to the real number space of $\mathrm{d} \times \mathrm{d}$. They are the learnable parameters of the three linear projection layers. SA represents the self-attention operation. $\mathrm{T}^{\mathrm{l}-1}$ represents the output of the $(\mathrm{l}-1)$-th layer. T represents the output of the l-th layer. d_{h} represents the dimension of tokens, which is also equal to d. It should be noted that the location of keypoints is typically predicted using heatmap [29–31].

4. Experiments

4.1. Experimental Details

4.1.1. Dataset

We employed DatPose for the COCO and MPII datasets [32]. The COCO dataset consists of more than 330 k images, 1.5 million targets and 80 target categories, and 91 material categories, and is publicly available. It has more than 250,000 keypoint marked pedestrians. The COCO dataset is usually used as an evaluation criterion for human pose estimation. MPII is a large-scale multi-person pose estimation dataset [21], which contains about 25,000 image samples. These images contain the poses of the characters in different

scenes and provide 16 keypoints of the characters, including the positions of keypoints such as head, torso, and limbs.

4.1.2. Evaluation Metrics

Following the metrics in [9], the standard average precision and recall rate are calculated to evaluate performance. In the COCO dataset, the performance of object keypoint detection models is evaluated using metrics such as average precision (AP) and average recall (AR). These metrics are calculated based on the object keypoint similarity (oks), which measures the similarity between predicted and ground truth keypoint locations:

$$\mathrm{OKS} = \sum_i \frac{\exp\left(-\frac{\mathrm{d_i^2}}{2\mathrm{s^2 k_i^2}}\right)\delta(\mathrm{v_i} > 0)}{\delta(\mathrm{v_i} > 0)} \tag{12}$$

where $\mathrm{d_i}$ represents the Euclidean distance between the i-th predicted keypoint coordinate and its corresponding ground truth. $\mathrm{v_i}$ represents the visibility flag of the keypoint. s denotes the object scale, and $\mathrm{k_i}$ is a constant specific to each keypoint.

The quantity of $\mathrm{d_i}$ in the given equation represents the Euclidean distance between the detected keypoint and the corresponding ground truth. The visibility flag of the ground truth is represented by $\mathrm{v_i}$. The object scale is denoted by s. Additionally, $\mathrm{k_i}$ is a per keypoint constant that governs the falloff rate. As such, this expression plays a significant role in assessing the efficacy of keypoint detection algorithms. The key point evaluation criterion of the MPII dataset is the head-normalized probability of correct keypoint (PCKh), and its formula is expressed as:

$$\mathrm{PCKh@}\alpha = \frac{\sum_{i=1}^{J} f(p_i)@\alpha}{\mathrm{X}} \tag{13}$$

Among them, PCKh@α is the proportion of keypoints correctly predicted when the head threshold is α, X is the number of keypoints, and f(p_i) is the similarity of i th keypoint.

4.1.3. Implementation Details

The experimental operating system is Ubuntu 18.04, the programming environment is PyTorch 1.10.1 + cu113, Python 3.8.12, and the GPU is NVIDIA Tesla T4. We increase the height or width of the human detection box to a predetermined aspect ratio: 4:3, and subsequently crop the box from the image, which is resized to a fixed dimension of either 256×192 or 384×288. The data augmentation techniques incorporated during this process comprise random rotation (within the range of -45° to 45°), random scaling (between 0.65 and 1.35), and flipping. In this work, we follow the two-stage top-down human pose estimation paradigm, which has been utilized in several prior works such as [5,17,33,34]. The approach involves initially detecting the individual person instance using a person detector and subsequently predicting the keypoints. To accomplish this, we adopt the popular person detectors furnished by SimpleBaseline [5] for both the validation set and test-dev set. The input image size is set to 256×192. The mean square error loss is used for learning. The Adam optimizer [35] was utilized to train our model for a total of 300 epochs. Throughout the training process, a small batch size of 16 and a dropout rate of 0.5 were employed. The initial learning rate is 1×10^{-3}. The predicted heatmaps are two-dimensional spatial information, and we use the two-dimensional sine strategy to embed the position. Figure 5 shows visual outcomes attained by the proposed DatPose model on MS COCO, which encompasses diverse scenarios. Our model has demonstrated precise prediction capabilities for various challenging scenarios such as variations in viewpoint and appearance, as well as instances of occlusion.

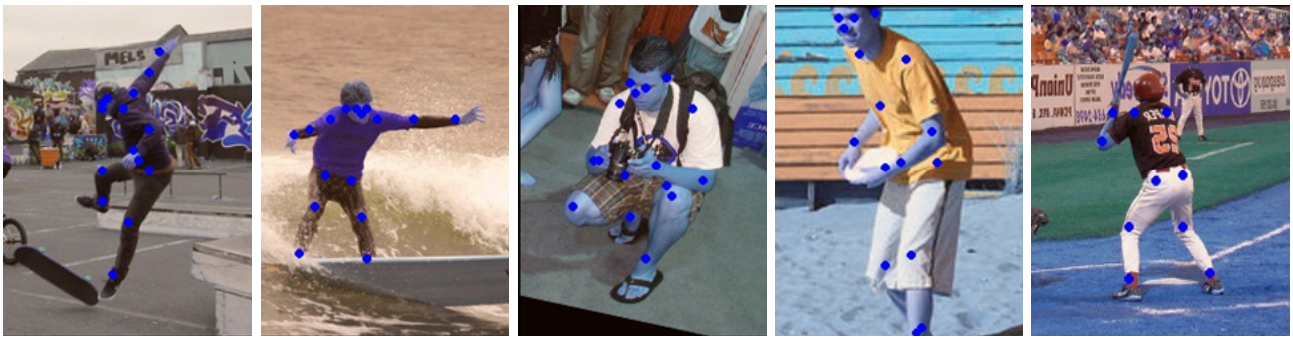

Figure 5. Qualitative results of some example images in the COCO data set containing view-point and appearance change, occlusion.

4.2. Comparison with State-of-the-Art Methods

Table 1 shows the comparison of DatPose with state-of-the art models, including the CNN-based methods [5,17] and CNN-based methods proposed spatial multiple scales features. The CNN-Transformer based methods [8,9,36] capture the constraints of spatial locations. The pure Transformer model learns the relationship between features directly from the original image [9,10]. Our model consistently outperforms state-of-the-art models on all the metrics and achieves 74.8% boost on AP and 80.3% boost on AR accuracy. Although the VITPose-B model improves the AP by 1% compared with the Datspose model, it is worth noting that the Datspose model has fewer parameters and reduces the complexity of the model.

Table 1. State-of-the-art comparison on COCO validation set.

Method	Pretrain	Input Size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
SimpleBaseline-Res50	Y	256 × 192	34.0 M	8.9	70.4	88.6	78.3	67.1	77.2	76.3
HRNet-W32	N	256 × 192	28.5 M	7.1	74.4	90.5	81.9	70.8	81.0	79.8
TokenPose-S-v1	N	256 × 192	6.6 M	2.2	72.5	89.3	79.7	68.8	79.6	78.0
TransPose-R-A4	Y	256 × 192	6.0 M	8.9	72.6	-	-	-	-	78.0
FET	N	256 × 192	8.2 M	5.9	72.9	-	-	-	-	78.1
VITPose-B	Y	256 × 192	86 M	17.1	75.8	90.7	83.2	-	-	81.1
DatPose	N	256 × 192	6.9 M	3.3	74.8	91.4	80.3	69.2	77.5	80.3

Table 2 compares the performance of this algorithm with other methods on the COCO test-dev set. Compared with HRNet, the AP is improved by 0.5%, indicating superior performance. Moreover, compared with HRNet, the Params and GFLOPs indexes of our method are significantly reduced, thus ensuring the lightweight of the model. Furthermore, when compared to TransPose [4], DatPose achieves the same AP while utilizing only 32% of TransPose's [4] GFLOPs. Compared with TokenPose [5], the AP is slightly inferior, but it has fewer parameters and capacity. The reason is that the fusion module efficiently fuses high-level semantic information and spatial location detail information, thus commanding less capacity. Based on the above experimental results, the method proposed in this work has fewer parameters and complexity compared with the large model net-work. In addition, compared with the lightweight network, the accuracy of human pose estimation is improved under the condition of adding a small number of parameters, and it has the ability to compare with the advanced model.

Table 3 presents the experimental results of our algorithm compared to other state-of-the-art methods for human pose estimation on the MPII validation set. The input image size for all methods is set to 256 × 256 pixels. Our algorithm demonstrates a PCKh@0.5 improvement of 2.8% and 1.8% compared to the traditional convolution networks SHN and SimpleBase-Res50, respectively. Furthermore, when compared to a Transformer-based human pose estimation model, specifically the baseline TokenPose, our algorithm achieves a modest improvement of 0.1%.

Table 2. State-of-the-art comparison on COCO test-dev set.

Model	AP/%	AP^{50}/%	AP^{75}/%	AP^{M}/%	AP^{L}/%	AR/%	Params/M	GFlops
CPN [2]	72.1	91.4	80.0	68.7	77.2	78.5	58.8	29.2
HRNet-W48 [3]	74.2	92.4	82.4	70.9	79.7	79.5	63.6	14.6
TransPose-H-A4 [4]	74.7	91.9	82.2	71.4	80.7	-	17.3	17.5
TransPose-H-A6 [4]	75.0	92.2	82.3	71.3	81.1	-	17.5	21.8
TokenPose-L/D6 [5]	74.9	92.1	82.5	71.7	81.1	80.2	27.5	11.0
DatPose	74.7	92.3	82.3	71.8	79.4	77.3	7.2	5.6

Table 3. State-of-the-art comparison on MPII dataset.

Model	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean@0.5
SHN [6]	96.5	96.0	88.4	83.5	87.1	83.5	78.3	87.5
SimpleBase-Res50 [7]	96.4	95.3	89.0	83.2	88.4	84.0	79.6	88.5
HRNet-W32 [3]	96.9	96.0	90.6	85.8	88.7	86.6	82.6	90.1
HRNet-W48 [3]	97.3	95.3	89.9	85.5	88.1	85.0	81.8	89.4
TokenPose-L/D6 [5]	97.1	95.9	91.0	85.8	89.5	86.1	82.7	90.2
CAPose-s2 [8]	97.0	95.8	90.9	86.2	89.5	86.9	83.4	90.4
DatPose	97.3	95.6	90.1	85.7	89.8	86.6	82.4	90.3

The COCO dataset is visualized using DatPose, where each column depicts the 17 keypoints and each row displays the prediction of the keypoints from varied viewpoints in Figure 6. The representation provides comprehensive insights into the accuracy of the keypoint predictions. TokenPose is the most relevant model to DatPose, as it strengthens the keypoint information to jointly assess all the patches in the self-attention. However, it introduces the keypoint features and image clues equally to all the Transformer Blocks without giving greater weight to the keypoint information. By collecting the keypoint information of human body edges via the Fusion of Convolution and Self-Attention Block, our model achieves remarkable improvement.

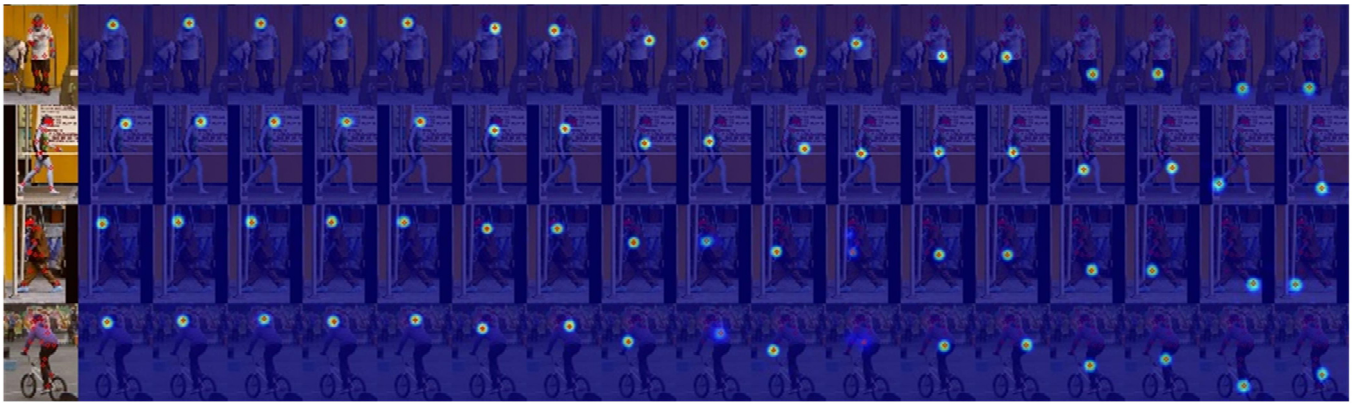

Figure 6. Visualization of DatPose on the COCO dataset. Each column represents the visualization of 17 keypoints, and each row represents the prediction of keypoints from different viewpoints.

4.3. Ablation Study

Table 4 shows ablation results to verify the contribution of each component in our model. Model '1' is a Transformer human pose estimation method based on the standard residual network ResNet. The models '2' and '3' are based on model '1', and the deformable convolution module and the fusion module are added, respectively, to compare the AP and AR.

Table 4. Ablation study of key components of DatPose.

Model	Deformable Convolution	Fusion Module	AP	AR
1	-	-	72.1	75.0
2	✓	-	72.5	78.0
3	-	✓	73.5	77.3
4	✓	✓	74.8	80.3

(1) In models '1' and '2', we evaluate the influence of the deformable convolution layer of the first stage on the performance. We observe that the AP of model '2' increases by 0.4% compared with model '1'. It indicates that the deformable convolution variant module is beneficial to grabbing the edge keypoint information of the human body. Therefore, the keypoint information has greater impact than the visual part.

(2) In models '1' and '3', we assess the effectiveness of the proposed fusion of convolution and self-attention module. Compared with model '1', the AP and AR of model '3' increased by 1.4% and 2.3%, respectively. By incorporating deformable convolutions and fusion modules into model 4, the AP and AR increased by 0.9% and 0.7%, respectively, compared to model '1'. In addition, the AP of model '3' is 1% higher than that of model '2', which shows that the fusion module has a great influence on the improvement of the AP of the full model. It proves that the keypoints are generated by the fusion module for the benefit of different visual weights when both models replace the deformable convolution and fusion module at the same time. Figure 7 depicts the relationship between keypoints and visual cues, where it is observed that the attention layer in the model results in an increased focus on fine-grained details, leading to a higher weighting of keypoints. Notably, the intensity of the red color corresponds to the significance of the information being considered.

(3) To verify the effect of the convolution of the fusion module, the ablation experiments were carried out, as seen in Table 5. The AP of the fusion of convolution and self-attention method is 1.6% and 1.4% higher than that of the only attention method and the keypoint with convolution method, respectively. It shows that the performance reduces in the case of only using the attention mechanism. It indicates that the feature extraction ability of the convolution was stronger than that of the attention mechanism.

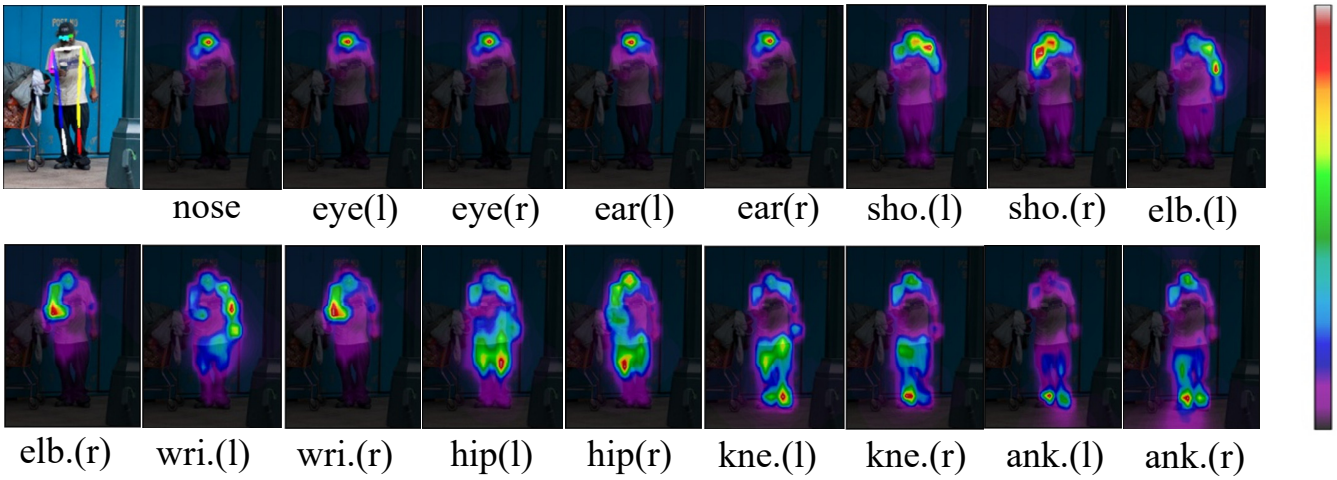

Figure 7. The visualization of attention maps based on the dependency relationship between keypoints and visual cues.

Table 5. Ablation study of fusion module on COCO dataset.

Method	AP	AR
Fusion of Convolution and Self-Attention	74.8	78.0
Only Self-Attention	73.2	77.8
Keypoint with convolution	73.4	77.5

5. Conclusions

In this paper, we propose a framework for human pose estimation named DatPose, which leverages external keypoint information and focuses on strengthening the weight of keypoints. We skillfully incorporate deformable convolution to capture human keypoints adaptively and introduce a convolution and attention fusion module to enhance the weight of key points in an image. Our model outperforms state-of-the-art methods with a smaller number of parameters and achieves interpretable results on benchmark datasets.

Author Contributions: Conceptualization, X.W., N.S. and G.W.; Data curation, X.W.; Formal analysis, J.S.; Funding acquisition, N.S.; Investigation, N.S.; Methodology, X.W.; Project administration, X.W.; Software, S.Z.; Supervision, J.S.; Validation, S.Z.; Visualization, S.Z.; Writing—original draft, X.W.; Writing—review and editing, X.W. and N.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Henan Province of China under Grant 232300420157.

Data Availability Statement: Publicly archived datasets used in the study are listed below. COCO: http://cocodataset.org (accessed on 12 May 2023); MPII: http://human-pose.mpi-inf.mpg.de (accessed on 8 June 2023).

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

References

1. Adama, D.A.; Lotfi, A.; Ranson, R. A Survey of Vision-Based Transfer Learning in Human Activity Recognition. *Electronics* **2021**, *10*, 2412. [CrossRef]
2. Mavrogiannis, P.; Maglogiannis, I. Amateur football analytics using computer vision. *Neural Comput. Appl.* **2022**, *34*, 19639–19654. [CrossRef]
3. Zhao, L.; Yang, F.; Bu, L.; Han, S.; Zhang, G.; Luo, Y. Driver behavior detection via adaptive spatial attention mechanism. *Adv. Eng. Inform.* **2021**, *48*, 101280. [CrossRef]
4. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part VIII 14; pp. 483–499. [CrossRef]
5. Xiao, B.; Wu, H.P.; Wei, Y.C. Simple Baselines for Human Pose Estimation and Tracking. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 472–487. [CrossRef]
6. Cheng, B.W.; Xiao, B.; Wang, J.D.; Shi, H.H.; Huang, T.S.; 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 (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 5385–5394. [CrossRef]
7. 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 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017. [CrossRef]
8. Yang, S.; Quan, Z.B.; Nie, M.; Yang, W.K. TransPose: Keypoint Localization via Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision Electr Network, Montreal, BC, Canada, 11–17 October 2021; pp. 11782–11792. [CrossRef]
9. Li, Y.J.; Zhang, S.K.; Wang, Z.C.; Yang, S.; Yang, W.K.; Xia, S.T.; Zhou, E.J. TokenPose: Learning Keypoint Tokens for Human Pose Estimation. In Proceedings of the 18th IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 11293–11302. [CrossRef]
10. Xu, Y.; Zhang, J.; Zhang, Q.; Tao, D. Vitpose: Simple vision transformer baselines for human pose estimation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 38571–38584. [CrossRef]
11. Yuan, Y.; Fu, R.; Huang, L.; Lin, W.; Zhang, C.; Chen, X.; Wang, J. Hrformer: High-resolution vision transformer for dense predict. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 7281–7293. [CrossRef]

12. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929. [CrossRef]
13. Liu, Z.; Lin, Y.T.; Cao, Y.; Hu, H.; Wei, Y.X.; Zhang, Z.; Lin, S.; Guo, B.N. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 18th IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002. [CrossRef]
14. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jegou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning (ICML), Electr Network, Virtual Event, 18–24 July 2021; pp. 7358–7367. [CrossRef]
15. Rao, Y.M.; Zhao, W.L.; Liu, B.L.; Lu, J.W.; Zhou, J.; Hsieh, C.J. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), Electr Network, Virtual Event, 6–14 December 2021. [CrossRef]
16. Jiang, M.X.; Yu, Z.L.; Li, C.H.; Lei, Y.Q. SDM3d: Shape decomposition of multiple geometric priors for 3D pose estimation. *Neural Comput. Appl.* **2021**, *33*, 2165–2181. [CrossRef]
17. Sun, K.; Xiao, B.; Liu, D.; Wang, J.D.; Soc, I.C. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 5686–5696. [CrossRef]
18. Wu, C.; Wei, X.; Li, S.; Zhan, A. MSTPose: Learning-Enriched Visual Information with Multi-Scale Transformers for Human Pose Estimation. *Electronics* **2023**, *12*, 3244. [CrossRef]
19. Wu, H.P.; Xiao, B.; Codella, N.; Liu, M.C.; Dai, X.Y.; Yuan, L.; Zhang, L. CvT: Introducing Convolutions to Vision Transformers. In Proceedings of the 18th IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 22–31. [CrossRef]
20. Dong, X.Y.; Bao, J.M.; Chen, D.D.; Zhang, W.M.; Yu, N.H.; Yuan, L.; Chen, D.; Guo, B.N.; Ieee Comp, S.O.C. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 12114–12124. [CrossRef]
21. Peng, Z.L.; Huang, W.; Gu, S.Z.; Xie, L.X.; Wang, Y.W.; Jiao, J.B.; Ye, Q.X. Conformer: Local Features Coupling Global Representations for Visual Recognition. In Proceedings of the 18th IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 357–366. [CrossRef]
22. Dai, J.F.; Qi, H.Z.; Xiong, Y.W.; Li, Y.; Zhang, G.D.; Hu, H.; Wei, Y.C. Deformable Convolutional Networks. In Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 764–773. [CrossRef]
23. Zhu, X.Z.; Hu, H.; Lin, S.; Dai, J.F.; Soc, I.C. Deformable ConvNets v2: More Deformable, Better Results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9300–9308. [CrossRef]
24. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Muller, K.R.; Samek, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE* **2015**, *10*, 46. [CrossRef] [PubMed]
25. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [CrossRef]
26. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv* **2013**, arXiv:1312.6034. [CrossRef]
27. Pan, X.R.; Ge, C.J.; Lu, R.; Song, S.J.; Chen, G.F.; Huang, Z.Y.; Huang, G.; Ieee Comp, S.O.C. On the Integration of Self-Attention and Convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 805–815. [CrossRef]
28. Ba, J.L.; Kiros, J.R.; Hinton, G. Layer normalization. *arXiv* **2016**, arXiv:1607.06450. [CrossRef]
29. Yang, W.; Ouyang, W.L.; Li, H.S.; Wang, X.G. End-to-End Learning of Deformable Mixture of Parts and Deep Convolutional Neural Networks for Human Pose Estimation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3073–3082. [CrossRef]
30. Chu, X.; Yang, W.; Ouyang, W.L.; Ma, C.; Yuille, A.L.; Wang, X.G. Multi-Context Attention for Human Pose Estimation. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5669–5678. [CrossRef]
31. Chu, X.; Ouyang, W.L.; Li, H.S.; Wang, X.G. Structured Feature Learning for Pose Estimation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4715–4723. [CrossRef]
32. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the 13th European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755. [CrossRef]
33. Chen, Y.L.; Wang, Z.C.; Peng, Y.X.; Zhang, Z.Q.; Yu, G.; Sun, J. Cascaded Pyramid Network for Multi-Person Pose Estimation. In Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7103–7112. [CrossRef]
34. Papandreou, G.; Zhu, T.; Kanazawa, N.; Toshev, A.; Tompson, J.; Bregler, C.; Murphy, K. Towards Accurate Multi-person Pose Estimation in the Wild. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3711–3719. [CrossRef]

35. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980. [CrossRef]
36. Wang, D.; Xie, W.J.; Cai, Y.C.; Liu, X.P. A Fast and Effective Transformer for Human Pose Estimation. *IEEE Signal Process. Lett.* **2022**, *29*, 992–996. [CrossRef]