EMGPose: An Efficient Multi-Granularity Representation for Human Pose Estimation

Guonan Deng[1], Shiyong Lan[1,†], Wenwu Wang[3], Yixin Qiao[1], Yao Li[1], Haohan Chen[1], Hongyu Yang[1,2]
[1]College of Computer Science, Sichuan University, Chengdu, 610065, China
[2]National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu, 610065, China
[3]Department of Electrical and Electronic Engineering, University of Surrey, Guildford, GU2 7XH, U.K.
dengguonan@stu.scu.edu.cn, lanshiyong@scu.edu.cn, w.wang@surrey.ac.uk, {Qiaoyixin, liyao518, sajel}@scu.edu.cn

***Abstract*—Current Transformer-based methods typically unwisely represent the entire image at a single granularity. A high-resolution representation of the regions of interest can significantly improve the accuracy of human pose estimation while causing unnecessary computational costs for other regions. To overcome this limitation, we propose an efficient two-stage framework using adaptive Multi-Granularity representation for different important image regions for Human pose estimation (EMGPose). In the first stage, the image is split into coarse-grained patches for simple inference. If without sufficient accuracy, important patches will be resplit into multiple finer-grained patches for the second stage of inference. Furthermore, we propose a new token-merge strategy based on token importance and similarity in Transformer, effectively reducing the computational load from low-information background patches. Extensive experiments demonstrate the excellent performance of the proposed method. Specifically, our model EMGPose-Base achieves 76.3 AP (+0.5 AP) and 62.2 AP (+2.6 AP) and higher efficiency than baseline ViTPose-Base on the COCO validation set and OCHuman test set, respectively.**

***Index Terms*—human pose estimation, multi-granularity, efficient, two-stage, token merge**

I. Introduction

Human pose estimation (HPE) is a fundamental task in computer vision. It aims at locating anatomical keypoints of the human body (e.g. head, shoulders, elbows, etc.) in an image. HPE has been extensively researched and lays the foundation for diverse downstream tasks, such as medical diagnosis [1] and action recognition [2].

In recent years, human pose estimation has made significant advancements with the predominant methods employing output heatmaps and subsequently utilizing the peaks of the heatmaps as keypoint locations. For instance, HRNet [3] and Lite-HRNet [4] achieve rich semantic information and precise localization by parallelizing multiple resolution branches of convolutional neural network (CNN). Moreover, Transformer and its variants are widely used in HPE tasks due to their powerful ability to model long-range dependencies. Representative TokenPose [5] and DTPose [6] regard keypoints as learnable representations, establishing a link between visual and keypoint tokens through a combination of CNN and Transformer, which enables the extraction of pose features with notable performance. Furthermore, ViTPose [7] suggests that competitive performance in pose estimation can be achieved without using an elaborate Transformer framework.

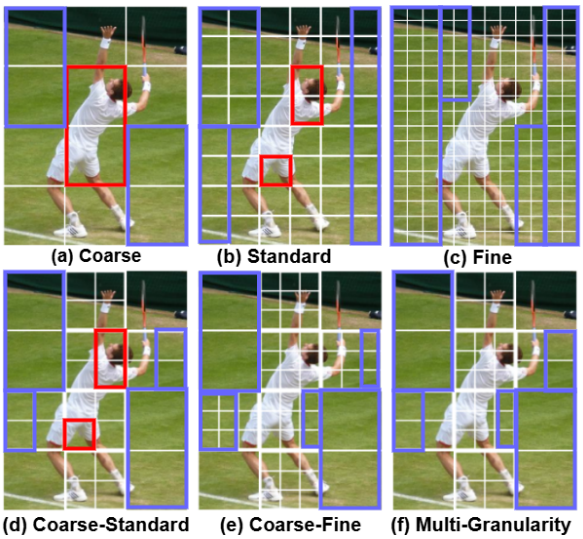

Fig. 1: Image splitting example graph for different strategies, where (a-c) are one-stage single granularity. (d-e) are two-stage single granularity. (f) is our proposed two-stage multi-granularity method. The **red boxes** in the figure indicate joints coupling and the **blue boxes** indicate background redundancy.

However, the localization of the body keypoints from an image is often challenging due to the variations in occlusion, truncation, scale, and human appearance. Prior research [8], [9] has shown that utilizing high-resolution representations can enhance regression accuracy but at the cost of greatly increased computation. This presents a significant challenge for Transformer-based models, where the computational complexity grows quadratically with the number of input patches. The number of patches is closely tied to the splitting strategy selected by the Transformer-based HPE model. Current methods [5]–[8] primarily use a single granularity for image splitting (as shown in Fig.1). From this, it can be analyzed that many image patches lack effective contextual information and vary in importance. Excessive coarse splitting (Fig.1(a),(d))

† Corresponding Author.
This work was supported by the National Natural Science Foundation of China project 62371324. This work was also supported by the 2035 Innovation Pilot Program of Sichuan University, China.
The code is available at https://github.com/SYLan2019/EMGPose.

can couple multiple joints in a single patch, which leads to the under-representation of important keypoints. Excessive fine splitting (Fig.1(c),(e)) renders the wastage of computational resources. Therefore, inspired by [10], considering the differences in information density of each image patch: Image patches should be characterized at varying granularity according to their keypoint information content.

Furthermore, Fig.1 illustrates that there are still numerous low-information background patches in either image-splitting method, which increase with finer granularity. To alleviate this burden and enhance efficiency, many works [11]–[16] focus on token compression. In common methods, token merge [11]–[14] typically merges tokens with high similarity on the basis of a similarity matrix. Meanwhile, token prune [15]–[17] identifies and discards unimportant tokens based on specific criteria, which results in the loss of detailed edge information and consequently poor performance. For example the token-Pruned Pose Transformer (PPT) [17] incorporates a human token identification module with attentional token pruning. SHaRPose [8] employs a two-stage framework that transitions from coarse to fine to speed up inference. However, it performs poorly on low-resolution images due to the limitations of single granularity segmentation. Although these methods accelerate Transformers, they greatly sacrifice accuracy.

Based on the above, we present EMGPose, a human pose estimation architecture for multi-granularity modeling based on the pure Transformer. This approach aims to improve prediction accuracy while optimizing the computational cost. To reduce the computational burden, the patches with strong correlations to body keypoints are re-split at varying granularity of importance. Moreover, the weakly correlated background patches mentioned above are gradually fused in the subsequent processing by our proposed merge module. Overall, the main contributions of this paper are as follows:

- EMGPose employs a two-stage multi-granularity HPE framework, which adaptively accounts for the differences in the importance of the image patches in pose estimation.
- We propose a Weighted token merge module Based on the synergy of Attention scores and Similarity matrices (WBAS), which compresses the tokens while maintaining decent accuracy.
- EMGPose shows superior performance while improving the efficiency of the pure Transformer model in the pose estimation task. On the universal MS COCO dataset, our proposed model outperforms the state-of-the-art methods.

II. Method

A. Overall Structure

As depicted in Fig.2, EMGPose contains two stages with a shared keypoint heatmap head and Transformer blocks. Each image is initially subjected to the Coarse-Granularity Stage (CGS), which incorporates a Transformer and a Quality Predictor. If the Predictor deems the result unacceptable, the process progresses to the Multi-Granularity Stage (MGS). The MGS encompasses a patch Re-Patchfy module, a Feature Reuse module, a WBAS module, and a Transformer that shares parameters with the Coarse-Granularity Stage.

In this section, we will present our framework stage-by-stage and give a detailed introduction to each module.

B. Coarse-Granularity Stage (CGS)

The objective of this stage is to identify keypoints in the coarse-grained image and ascertain their acceptablity for the Quality Predictor. For challenging pose input images, the importance scores obtained during the CGS process are used to efficiently select the important patches for the next stage.

Coarse-grained Input. Firstly for efficient coarse-grained inference, we use the downsampled copy $\mathrm{X}' \in \mathbb{R}^{(H/2\times W/2\times C)}$ of the input image $\mathrm{X} \in \mathrm{R}^{(H\times W\times C)}$ for simple inference. Similar to standard Vision Transformers (ViT), which utilize fixed-size 2D patches for segmentation. With the addition of auxiliary information tokens, the token input about the CSG can be represented as follows:

$$I_0^C = Patchfy(Downsample(X)) + E_{pos}^C \in R^{N\times D} \quad (1)$$

$$X_0^C = Concat(K_0^C, I_0^C, q_0) \in R^{(T+N+1)\times D} \quad (2)$$

where N represents the number of image patches and C stands for CGS. The patches are subsequently mapped into D-dimension tokens by patch embedding. Additionally, position embedding E_{pos}^C is included to enrich position information. T learnable D-dimensional embedding vectors $K_0^C = \{k_0^{i,C}\}_{i=1}^T$ are employed to represent T body keypoints. Additionally, a learnable quality token q_0 is utilized to estimate confidence results. An L-layer ViT then processes the input sequence to obtain the output sequence $X_L^C = [k_L^{1,C}, k_L^{2,C}, ..., k_L^{T,C}, i_L^{1,C}, i_L^{2,C}, ..., i_L^{N,C}, q_L]$, each layer comprises a Multi-Head Self-Attention (MHSA) module and a Feed-Forward Network (FFN). We can represent the output X_{l+1} of layer $l+1$ as follows:

$$X_l^{'} = MHSA(X_l) + X_l, \quad X_{l+1} = FFN(X_l^{'}) + X_l^{'}. \quad (3)$$

To obtain the result of CGS inference, we take out the corresponding keypoint tokens $K_L^C = \{k_L^{i,C}\}_{i=1}^T$ from the output sequence X_L^C that has undergone L-layer Vision Transformer inference and put them into the Heatmap Head consisting of Multilayer Perceptron (MLP):

$$R^C = Head(K_L^C) \in R^{M\times \hat{H}\times \hat{W}}. \quad (4)$$

Here R^C represents the regression heatmap results obtained at the CGS stage. The size of the heatmap, represented by $\hat{H}\times\hat{W}$, is equivalent to one-quarter of the original input image size H and W.

Regression quality judgement. Our method used to evaluate the quality of the regression outcome follows the classical two-stage approach [8], [18]. That is, leveraging a learnable quality embedding q_0 to extract information from visual and keypoint tokens during the inference, resulting in reasonable quality features q_L. Subsequently, q_L was fed into the Quality Predictor module, which produced the quality score Q. For simplicity and efficiency, we use MLP as this predictor thus

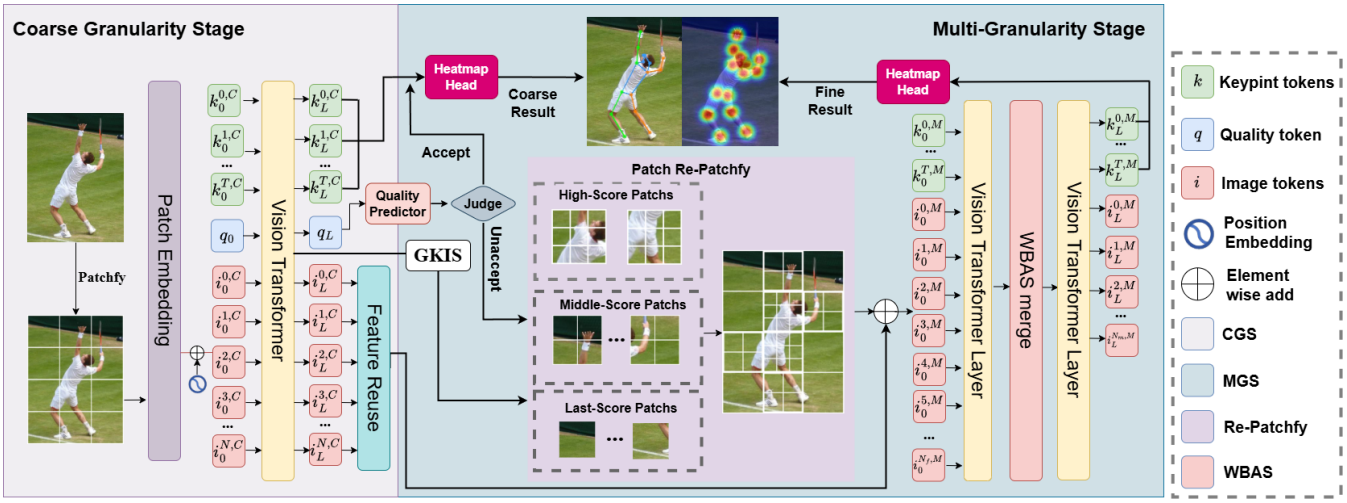

Fig. 2: The overall structure of **EMGPose**. The **Global Keypoint Importance Score (GKIS)** yielded by the Transformer in the **Coarse-Granularity stage** is used for selecting keypoint-related important patches in the **Multi-Granularity stage** to **re-patchfy** the image. The parameters of the Transformer blocks and the Heatmap Head are **shared** between the two stages.

obtaining the quality score $Q = MLP(q_L)$. We introduce a threshold η as a criterion for deciding whether to accept the coarse result R^C. If $Q > \eta$, the inference process ends immediately, and R^C becomes the final prediction output. Otherwise, the input image must undergo further estimation by entering the Multi-Granularity Stage. Notably, η strikes a balance between performance and computational efficiency.

C. Multi-Granularity Stage (MGS)

Higher resolution (i.e., Finer granularity) usually leads to higher accuracy. However, since computational resources are limited, we only use fine-grained representation at important locations. We use the attention computed in the previous stage as an importance score (IS) that reflects the correlation between image patches and keypoints without extra computation.

$$IS^l = \frac{1}{HT}\sum_{i=1}^{T}\sum_{\mathrm{h}=1}^{H} A^l_{i,\mathrm{h}} \in R^N, \tag{5}$$

where $A^l_{i,\mathrm{h}}$ represents the attention vector of the i-th keypoint token from the l-th layer of ViT with the image tokens at the h-th head. H and T indicate the number of heads of MHSA and the number of keypoints, respectively. To represent token importance more accurately and stably, we use the exponential moving average (EMA) to combine the attention score of Transformer layers:

$$\hat{IS}^l = \beta \cdot \hat{IS}^{l-1} + (1-\beta)IS^l, \tag{6}$$

where β = 0.98 and attention is unstable at shallow layers, so calculations are averaged over EMA from $l > 3$ onwards. The final layer's $\hat{IS}^L$ is used as the Global Keypoints Importance Score (GKIS) to discriminate patches.

Re-Patchfy. As illustrated in Fig.2, image tokens are categorized into three levels of importance according to the GKIS: high, middle, and low. Each category is numbered as:

$$N_h = \lfloor N \cdot \mathbf{r}_h \rfloor, N_l = \lfloor N \cdot \mathbf{r}_l \rfloor, N_m = N - N_h - N_l, \tag{7}$$

where r_h and r_l represent the rate of high-score and low-score patches, respectively. Given that the high-score patches contain the most crucial keypoint information, each of them will be subdivided into 3×3 fine-granularity patches. The medium-scored patches have the second most important keypoint information, and each patch will be subdivided into 2×2 strand-granularity patches. Conversely, the low-scored patches will retain their original coarse-granularity. Re-Patchfy image token sequence is represented as: $I_0^M = [\hat{i}_0^{1,M}, \hat{i}_0^{2,M}, ..., \hat{i}_0^{N_f,M}] \in R^{N_f \times D}$, where $N_f = 9 \times N_h + 4 \times N_m + N_l$ represents the number of image patches after repartitioning and M stands for MGS.

Feature Reuse. Multi-granularity splitting weakens the correlation between image patches. To address this limitation, We reuse the image feature $I_L^C = \{i_L^{j,C}\}_{j=1}^N$ extracted from the CGS output sequence X_L^C, which contains rich global correlation information. The feature I_L^C is first processed by the MLP layer to enhance the interconnection between the features. These feature tokens were then interpolated to 4 × and 9 × to align the re-pathfied image token sequence I_0^{M} following the same processing as Re-Pathfy. Denote as:

$$X^{FR} = FR(I_L^C) = [\mathrm{i}_{fr}^1, \mathrm{i}_{fr}^2, ..., \mathrm{i}_{fr}^{N_f}]. \tag{8}$$

Ultimately, the input sequence of the Multi-Granularity Stage can be represented as:

$$X_0^M = Con(K_0^M, (I^M + X^{FR} + E_{\mathrm{pos}}^{\mathrm{M}})) \in R^{(T+N_f)\times D}, \tag{9}$$

where Con stands for Concat operate and $K_0^M = \{k_0^{i,M}\}_{\mathrm{i}=1}^T$ is the same initial keypoint embedding as in (2). E_{pos}^M represents the positional embedding of MGS.

WBAS. The multi-granularity representation still has many low-information tokens resulting in the computational burden. Token prune usually loses image information and token merge only considers the similarity that may result in the important tokens being merged, both of which will lead to a significant

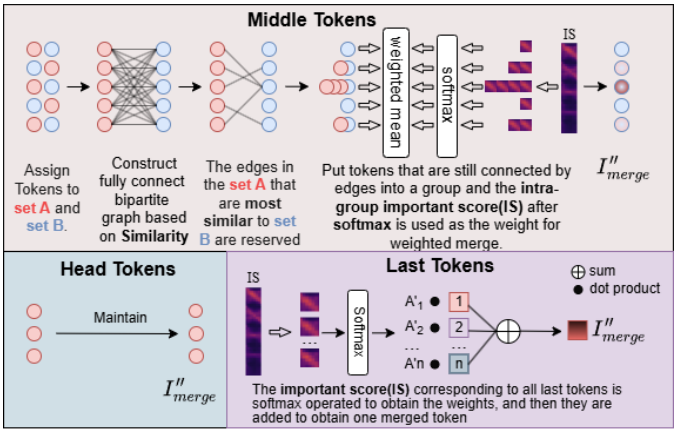

Fig. 3: **WBAS** Detail Schematic. WBAS adopts the **three-way decision** for Head, Middle, and Last tokens to adopt different suitable strategies to deal with them.

TABLE I: Configurations of the EMGPose models.

Method	Feat. Dim.	Depths	r_h/r_l	r_H/r_L	Merge Layer
EMGPose-Small	384	12	0.1/0.4	0.5/0.1	[4 , 7 , 10]
EMGPose-Base	768	12	0.1/0.4	0.5/0.1	[4 , 7 , 10]

reduction in accuracy. To address this issue, we refine them further to compress tokens while maintaining decent accuracy.

Unlike earlier merge methods, our approach builds on our previous idea that different tokens with varying importance should be handled distinctly. Therefore, we also use three-way decisions to make different merge strategies for tokens with different importance levels. The importance judgment uses (5) to calculate the importance score (IS) of the current layer. The tokens are classified into three categories according to the IS: Head, Middle, and Last. The number of tokens in each category is controlled by setting the ratio r_H and r_L of the Head and Last tokens. As shown in Fig.3, the Head tokens remain unchanged. The optimization of the Middle tokens $T_{\text{mid}} = \{t_i\}_{i=1}^{n}$ is achieved according to the following rules:

- Divide Middle tokens into two sets: $A = \{t_1, t_3, ..., t_{n-1}\}$ and $B = \{t_2, t_4, ..., t_n\}$.
- A **fully connected bipartite graph** is constructed between the tokens in two sets based on **cosine similarity**.
- Retain edges from a token in set A to the token in set B that indicates **the highest similarity** only.
- Tokens that remain connected by edges are considered as a group, and the **importance scores after softmax** are used as **weights** to average tokens **within the group**.

For Last tokens, the importance scores are directly weighted using softmax values to merge them into a single token.

We use the WBAS module for layers 4, 7, 10 of ViT in the Multi-Granularity Stage to implement token compression. The final output sequence is obtained as X_L^M. Finally, the keypoint tokens K_L^M are fed into the shared heatmap head defined in (4) to obtain finer inferred heatmap results $R^{\text{M}} = Head(K_L^M)$.

Training Strategies. The training objective for EMGPose is to supervise two stages of output heatmaps and quality predictors. First, for the heatmap, we employ the mean squared error loss (MSE), represented as follows:

$$L_h(R) = \frac{1}{T}\sum_{\text{i}}^{T} L_{MSE}(R^i, H_{gt}^i), \tag{10}$$

where H_{gt}^i represents the ground-truth heatmap of the i-th keypoint. Furthermore, Object Keypoint Similarity (OKS) [19] functions as a common metric for assessing the accuracy (i.e., confidence) of human keypoint results against the ground truth. Therefore, to enhance the reliability of the quality predictor, we apply an L2-norm loss between the quality score Q and the OKS calculated with the results of the Coarse-Granularity Stage R^C. In summary, the total loss function is as follows:

$$L = L_h(R^C) + L_h(R^M) + \lambda \parallel Q - OKS(R^C) \parallel_2 . \tag{11}$$

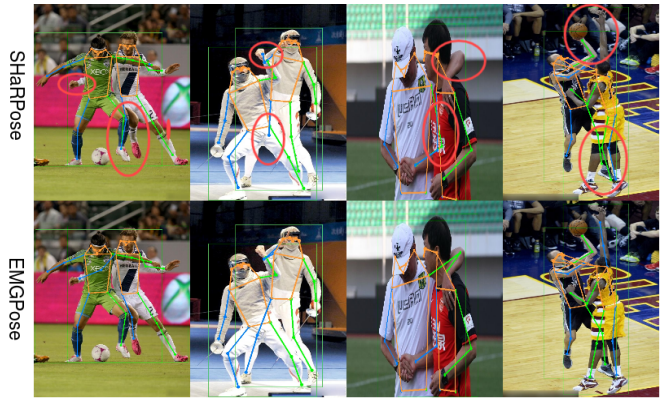

Fig. 4: illustrative figure of keypoint results for EMGPose-Base and SHaRPose-Base on the OCHuman dataset.

III. Experiments

A. Experiments Setup

Dataset. We evaluated the performance of EMGPose on the MS COCO [21] and OCHuman [22] datasets. we utilize the COCO 2017 dataset, which consists of more than 200k images and 250k human samples labeled with 17 keypoints. The dataset is divided into 3 subsets: train, valid, and test-dev, which contains 150k, 5k, and 20k samples respectively. Moreover, the OCHuman dataset, designed to solve the problem of high occlusion in human images, contains 5081 images totaling 13,360 human instances, making it the most complex and challenging dataset related to humans. We use the standard Average Precision (AP), Average Recall (AR), and Throughput as the main evaluation metric to assess the model performance.

Implementation Details. In this paper, all our experiments use a 16×16 patch size for splitting the images. We instantiate EMGPose with two different sizes by scaling the embedding size. The detailed configurations of the instantiated EMGPose models are presented in TableI. In the course of EMGPose training, we always set the confidence threshold $\eta = 1$, which implies that all images need to be inferred by MGS. Following [8], we set $\lambda = 0$ in the first 180 epochs and $\lambda = 0.03$ in the subsequent epochs. To ensure a fair comparison, all

TABLE II: Comparisons on the COCO valid and test-dev sets. No extra training data is involved for all results. The Throughput of all methods is recorded on a single RTX4090 GPU with a batch size of 64. The best result is highlighted in bold.

Model	Input	COCO val2017↑				COCO test-dev2017↑				Throughput↑	GLOPS↓
		AP	AP^L	AP^M	AR	AP	AP^L	AP^M	AR		
Dtpose-T [6]	256×192	69.4	75.5	66.6	75.3	68.9	75.0	65.6	73.8	863.6	**2.2**
ViTPose-Small [7]	256×192	73.8	75.8	67.1	79.1	73.1	78.5	70.1	78.5	1088.3	5.7
SHaRPose-Smalll [8]	256×192	74.2	80.3	71.2	79.5	73.6	79.0	70.7	79.0	**1290.6**	4.9
EMGPose-Small	256×192	**74.6**	**81.1**	**71.6**	**79.9**	**73.8**	**79.3**	**70.9**	**79.2**	1251.9	4.6
HRNet-W48 [3]	256×192	75.1	81.8	71.5	80.4	74.6	80.3	71.2	79.9	598.7	15.8
HRFormer-Base [20]	256×192	75.6	82.6	71.7	80.8	74.5	80.3	71.1	79.8	180.6	13.8
TokenPose-L/D24 [5]	256×192	75.8	82.7	72.3	80.9	75.1	**81.1**	71.7	80.2	456.9	11.0
PPT-L/D6 [17]	256×192	75.2	82.4	71.7	80.4	74.3	80.6	71.2	79.6	669.4	**9.2**
Dtpose-B [6]	256×192	75.7	82.8	71.9	80.7	74.8	80.8	71.2	79.8	533.8	10.6
ViTPose-Base [7]	256×192	75.8	78.4	68.7	81.1	75.1	80.7	72.0	80.3	614.6	18.6
SHaRPose-Base [8]	256×192	75.5	82.2	72.2	80.8	74.5	80.2	71.2	79.8	**712.8**	17.1
EMGPose-Base	256×192	**76.3**	**83.1**	**72.5**	**81.3**	**75.4**	**81.1**	**72.1**	**80.5**	688.8	17.1

TABLE III: Quantitative results from the OCHuman test set.

Method	Input	AP ↑	AP^{50} ↑	AR ↑	AR^{50} ↑
HRNet-W48 [5]	384×288	61.6	74.9	65.3	77.3
HRFormer-B [6]	384×288	49.7	71.6	58.2	76.0
ViTPose-Base [7]	256×192	59.6	74.7	64.1	77.8
SHaRPose-Base [8]	256×192	60.2	76.8	64.5	79.3
EMGPose-Base	256×192	**62.2**	**78.1**	**66.6**	**80.7**

TABLE IV: Comparison of different Splitting strategies.

Splitting Strategy	AP	AR	Throughput	GFLOPs
Coarse	52.0	66.2	2018.4	1.2
Standard	73.8	79.1	1088.3	5.7
Fine	75.7	80.9	460.3	13.4
Coarse-Standard	74.1	79.5	1290.6	4.9
Coarse-Fine	75.5	80.9	862.2	7.3
Coarse-Multi	75.4	80.7	1104	5.6

TABLE V: Comparison with efficient Transformers.

Method	AP	AR	Throughput	GFLOPs
Original	74.0	79.4	1290.6	4.9
EViT [15]	71.4	77.0	1438.5	3.52
DynamicViT [16]	69.2	75.1	824.6	3.52
ToMe [11]	71.6	77.3	1430.2	3.48
LF-ViT [18]	72.5	78.3	1282.3	4.51
WBAS(w/o[a] weighted merge)	72.8	78.6	1482.4	3.62
WBAS(w/o three-way decisions)	71.8	77.6	1389.6	3.48
WBAS	73.5	79.0	1441.2	3.85

[a]w/o stands for without

experiments presented in this paper are conducted using the MMPose [23] framework and the default data pipelines. Other optimal settings are set the same to ViTPose. To explore the potential of a pure Transformer model, both MAE [24] pre-training weights and Unbiased Data Processing (UDP) [25] post-processing are used in our approach.

B. Results

Comparison to state-of-the-art methods on MS COCO. To demonstrate the effectiveness of our dynamic framework in Human Pose Estimation(HPE). Table II shows the performance and efficiency of our proposed method with several state-of-the-art HPE methods on COCO valid and test-dev sets. We can see that we have achieved significant performance gains. For example, our EMGPose-Small model achieves 74.6 AP (+0.8 AP) and 73.8 (+0.7 AP) over ViTPose-Small in two sets respectively, and nearly 1.2x higher throughput, outperforming other methods in all accuracy metrics. Similarly, our EMGPose-Base model achieves 76.3 AP on the valid set. Notably, compared to the previous efficient models in HPE, SHaRPose-Base and PPT-L/D6, our model increases by a minimum of 0.8 AP at less than a 4% reduction in speed only. Our model also demonstrates faster inference speed than ViTPose-Base, TokenPose-L/D24, HRFormer-Base, HRNet-W48, and Dtpose-B, with superior accuracy. In addition, It can be observed that although the GFLOPS of EMGPose is not lower than that of Tokenpose, PPT, and other frameworks based on the combination of CNN and Transformer, it obtains a better trade-off between throughput and accuracy, showing that the pure Vision Transformer framework has strong representation ability and is friendly to modern hardware.

Comparison to state-of-the-art methods on OCHuman. We migrate the models with MS COCO data for training to test on the OCHuman dataset. The quantitative results on the test set are displayed in Table III. Our EMGPose performance is superior to other methods, even to the higher resolution of HRNet and HRFormer. For this significant enhancement we selected one of the outstanding single-grained methods, SHaRPose, for visualization and analysis. As shown in Fig.4, which shows that our EMGPose mitigates joints coupling (e.g. the part marked by **the red circle** in the figure) and achieves more accurate keypoint localization in the face of dense and occluded keypoint images compared to SHaRPose.

C. Ablation Study

Splitting Strategy. Choosing an appropriate granularity for image splitting is crucial in HPE as it affects accuracy and computation complexity. To verify the effectiveness of multi-granularity splitting, we use the ViTPose-Small as the baseline model and use the one-stage, two-stage single-granularity, and our multi-granularity strategies sequentially. the results are shown in Table IV, where Standard granularity refers to the 256×192 resolution image. Coarse and Fine represent its 1/2

TABLE VI: The effect of r at different settings

r_h / r_l	AP	GFLOPS	r_H / r_L	AP	GFLOPS
0.1 / 0.5	74.3	4.33	0.5 / 0.15	74.3	4.48
0.15 / 0.5	74.3	4.48	0.5 / 0.1	74.6	4.56
0.1 / 0.4	74.6	4.56	0.6 / 0.15	74.5	4.64
0.15 / 0.4	74.7	4.72	0.6 / 0.1	75.0	4.72

and 3/2 times resolution feature representation respectively. The two-stage approach offers a superior balance between accuracy and efficiency. Although fine-grained splitting attains optimal accuracy, it is remarkably resource-intensive. Consequently, our Multi-granularity strategy, which entails a 0.3 AP reduction in exchange for a 2.4x speed enhancement compared to fine-grained, is no doubt a rational choice.

Token Compression Policy. To demonstrate the superior performance of our proposed WBAS module, we compare it with several token compression policies. In addition, we performed ablation experiments without three-way decision and weighted merge to investigate the effect of our proposed fusion method. As shown in Table V, although our GFLOPS are higher than EViT and ToMe, we are not inferior in terms of real-world throughput and demonstrate that our method performs best in maintaining accuracy.

Ratio r. The two pairs of hyperparameters r are essential for managing the sparsity of the multi-granularity representation and the strength of redundant token fusion, which impacts the performance of EMGPose. r_h / r_l controls the number of tokens in multi-granularity representation. Most important information exists in a few high-score patches, a little in low-score patches. r_H / r_L controls the number of Head and Last tokens in the merge phase. We selected several competing control groups, the results of which are shown in Table VI. Above all, we set $r_h/r_t = 0.1/0.4$ and $r_H/r_L = 0.5/0.1$ respectively to match the speed-accuracy trade-off.

D. Conclusion

In this paper, we have proposed EMGPose, a two-stage efficient pose estimation framework that achieves a well-balanced trade-off between performance and computational cost. With Re-Pathfy, we strategically utilize varying levels of feature representation granularity for patches of differing importance. In addition, our proposed WBAS introduces a three-way decision mechanism and incorporates a weighted fusion of attention and similarity to achieve token compression applicable to the pose estimation task. Our quantitative experiments demonstrate the high accuracy and efficiency of our model. This work presents a pathway for investigating an efficient Transformer-based pose estimation framework.

REFERENCES

[1] Grzegorz Sarapata, Yuriy Dushin, Gareth Morinan, et al., "Video-based activity recognition for automated motor assessment of parkinson's disease," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[2] Zhen Xing, Qi Dai, et al., "Svformer: Semi-supervised video transformer for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18816–18826.

[3] Ke Sun, Bin Xiao, et al., "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[4] Changqian Yu, Bin Xiao, et al., "Lite-hrnet: A lightweight high-resolution network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021, pp. 10440–10450.

[5] Yanjie Li, Shoukui Zhang, et al., "Tokenpose: Learning keypoint tokens for human pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11313–11322.

[6] Shiyang Ye, Yuan Fang, Hong Liu, et al., "Dtpose: Learning disentangled token representation for effective human pose estimation," in *2024 IEEE International Conference on Image Processing*. IEEE, 2024, pp. 1336–1342.

[7] Yufei Xu, Jing Zhang, Qiming Zhang, et al., "Vitpose: Simple vision transformer baselines for human pose estimation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38571–38584, 2022.

[8] Xiaoqi An, Lin Zhao, et al., "Sharpose: Sparse high-resolution representation for human pose estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 691–699.

[9] Tsung-Yi Lin, Piotr Dollár, et al., "Feature pyramid networks for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[10] Yu Zhang, Yepeng Liu, Duoqian Miao, Qi Zhang, Yiwei Shi, and Liang Hu, "Mg-vit: a multi-granularity method for compact and efficient vision transformers," *Advances in Neural Information Processing Systems*, vol. 36, pp. 69328–69347, 2023.

[11] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, and Judy others, "Token merging: Your vit but faster," in *The Eleventh International Conference on Learning Representations*.

[12] Narges Norouzi, Svetlana Orlova, Daan de Geus, and Gijs Dubbelman, "Algm: Adaptive local-then-global token merging for efficient semantic segmentation with plain vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15773–15782.

[13] Daniel Bolya and Judy Hoffman, "Token merging for fast stable diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, June 2023, pp. 4599–4603.

[14] Zhanzhou Feng and Shiliang Zhang, "Efficient vision transformer via token merger," *IEEE Transactions on Image Processing*, 2023.

[15] Youwei Liang, Chongjian Ge, Zhan Tong, et al., "Not all patches are what you need: Expediting vision transformers via token reorganizations," in *International Conference on Learning Representations*, 2022.

[16] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh, "Dynamicvit: Efficient vision transformers with dynamic token sparsification," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13937–13949, 2021.

[17] Haoyu Ma, Zhe Wang, Yifei Chen, et al., "Ppt: token-pruned pose transformer for monocular and multi-view human pose estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 424–442.

[18] Youbing Hu, Yun Cheng, Anqi Lu, Zhiqiang Cao, Dawei Wei, Jie Liu, and Zhijun Li, "Lf-vit: Reducing spatial redundancy in vision transformer for efficient image recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 2274–2284.

[19] Sun Xiao, Xiao Bin, Wei Fangyin, Liang Shuang, and Yichen Wei, "Integral human pose regression," 2017.

[20] Yuhui Yuan, Rao Fu, Lang Huang, et al., "Hrformer: High-resolution vision transformer for dense predict," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7281–7293, 2021.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al., "Microsoft coco: Common objects in context," 2015.

[22] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul L. Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Hao-Zhi Huang, and Shi-Min Hu, "Pose2seg: Detection free human instance segmentation," 2019.

[23] MMPose Contributors, "Openmmlab pose estimation toolbox and benchmark," https://github.com/open-mmlab/mmpose, 2020.

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16000–16009.

[25] Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang, "The devil is in the details: Delving into unbiased data processing for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5700–5709.