Enhancing Human Pose Estimation with SE-Block in the OmniPose Model

Khac-Anh Phu
Faculty of Information Technology,
University of Sciences, Hue University.
Faculty of Information Technology,
Cao Thang Technical College
Hue City, Vietnam.
Ho Chi Minh City, Việt Nam
pkanh.dhkh23@hueuni.edu.vn

Van-Tuong-Lan Le[*]
University of Sciences, Hue University
Hue City, Vietnam
Corresponding to lvtlan@husc.edu.vn

Van-Dung Hoang[*]
Faculty of Information Technology,
HCMC University of Technology and Education
Ho Chi Minh City, Việt Nam
Corresponding to dunghv@hcmute.edu.vn

Thinh Vinh Le
Faculty of Information Technology,
HCMC University of Technology and Education
Ho Chi Minh City, Việt Nam
thinhlv@hcmute.edu.vn

2024 16th International Conference on Human System Interaction (HSI) | 979-8-3503-6291-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/HSI61632.2024.10613579

***Abstract*— The interaction and communication between humans and computers have brought diversity and richness to the field of computer vision research, offering numerous potentials and challenges in developing human action recognition applications. In this domain, recognizing human actions from image or video data plays a crucial role in various practical applications, from security surveillance to interactive control. Although the OmniPose model has demonstrated its effectiveness, there is still potential to improve its performance. In the scope of this study, we focus on enhancing the OmniPose model, which is used for extracting skeleton data from input image data. We propose two improvement methods: utilizing the Self-Attention mechanism and employing Squeeze-and-Excitation to enhance the skeleton data extraction capability of the OmniPose model. Through this approach, we aim to contribute to enhancing the performance of the OmniPose model in skeleton data extraction and human pose recognition, while opening doors to advancements in human action recognition in computer vision.**

Keywords*— *Computer vision, Deep learning, Human pose, Skeleton data.

I. Introduction

The task of skeleton data extraction plays a crucial role in the fields of computer vision and artificial intelligence. It focuses on identifying and reconstructing the structure of humans from image or video data by determining the positions and relationships between key points on the body, such as joints and other structural points. Thus, skeleton data not only highlights the position and movement of humans in space but also provides essential information for understanding their actions and behaviors [1].

The applications of skeleton data extraction are diverse and extensive. To date, it has been widely applied in various fields such as human-computer interaction, action recognition, animation, healthcare, sports, security, and many other applications. In human-computer interaction, skeleton data is used to recognize and respond to user actions, ranging from controlling devices to interacting with applications and user interfaces [2,3].

In the healthcare field, skeleton data is applied to monitor and evaluate physical exercises, facilitate functional recovery after injuries, and support treatment and rehabilitation processes. It can also be utilized in medical education and fitness training to provide feedback and effective exercise guidance. Additionally, in sports, skeleton data helps track and analyze the techniques, performance, and movements of athletes, thereby enhancing training and competitive capabilities.

Methods for extracting skeleton data and estimating 2D human poses have been extensively studied, with the emergence of many notable works such as [4, 5, 6, 7, 8]. Additionally, numerous research efforts have focused on developing methods for extracting skeleton data and estimating 3D human poses, as seen in studies like [9, 10,11]. Interest in this problem extends beyond processing data from a single individual [6], to include processing data from multiple individuals simultaneously [12]. Dealing with multiple individuals presents unique challenges for skeleton data extraction and pose estimation, particularly due to the large mechanical flexibility of the body and the simultaneous presence of overlapping joints. To address these challenges, approaches often rely on digital and geometric models to predict the positions of occluded joints [13, 14]. Among the various feature enhancement methods available, we chose the Squeeze-and-Excitation (SE) block for its significant ability to enhance important features and minimize irrelevant ones. Although SE-Block is not an attention mechanism in the traditional sense, it has been proven to significantly improve the performance of convolutional neural networks (CNNs) in various applications. The simplicity and computational efficiency of SE-Block make it an ideal choice for integration into the OmniPose model.

Furthermore, while attention mechanisms such as channel attention, spatial attention, and self-attention have demonstrated considerable success in many computer vision tasks [20], SE-Block provides a simpler and less computationally expensive alternative. Specifically, attention mechanisms often involve complex operations such as dynamic weight adjustment and self-attention, which, although powerful, can add significant computational overhead. In contrast, SE-Block achieves similar feature recalibration through a more straightforward and efficient approach, making it well-suited for our pose estimation tasks. The choice of SE-Block in this work reflects a balance between improving model performance and maintaining computational efficiency.

This paper contributes to the field of human pose estimation by improving the OmniPose model to enhance its understanding and prediction capabilities during data processing. Specifically, the main contributions of the paper are as follows:

- The paper proposes an enhancement method in the data preprocessing stage for the OmniPose model, specifically utilizing the SE-Block (Squeeze-and-Excitation Block). This integration opens up a new direction for improving the OmniPose model, particularly in skeleton data extraction and human pose estimation.

- Conducting experiments on the Max Planck Institute for Informatics Human Pose Dataset (MPII) dataset and evaluating the experimental results using the PCK index, the paper has demonstrated a significant improvement in the accuracy of extracting some important skeletal joints of the OmniPose model. This result is evidence that integrating SE-Block into the OmniPose model has brought about certain improvements in skeleton data extraction and human pose recognition tasks.

II. RELATED WORKS

A. HRNet

The High-Resolution Network (HRNet) has been widely used in various computer vision applications where high accuracy is required for identifying and localizing key points in images. One common application is human pose estimation, where HRNet excels in capturing detailed information at multiple resolutions, aiding in the precise identification of body joints and tracking body movements [8]. The HRNet model connects subnetworks from high to low resolutions in parallel, maintaining high-resolution heatmap predictions. This helps enhance the spatial accuracy of the heatmaps. HRNet also performs repeated multi-scale fusion to improve high-resolution representation, enriching the model with information. Experiments on COCO and MPII datasets have demonstrated the effectiveness of HRNet in human pose estimation and skeleton data extraction [15, 16]. This makes it suitable for applications where determining the specific location of image information is crucial, such as in medical imaging, autonomous driving systems, and augmented reality applications. Fig. 1 illustrates the architecture of the HRNet model.

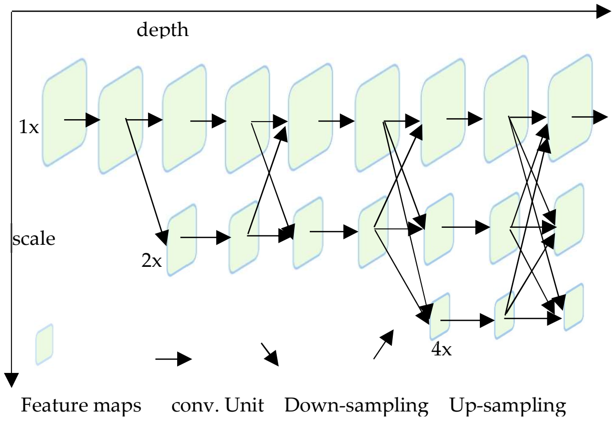

Fig. 1. HRNet Model

The HRNet model maintains high-resolution representations throughout the entire process, from input to output. The network comprises multiple branches with different resolutions operating in parallel, continuously integrating information through information exchange blocks. Consider Nsr as the subnetwork at stage s, and r as the resolution index. The following presents an example of a network structure that includes four parallel subnetworks.

$$\begin{tikzcd}
N_{11} \arrow[r] \arrow[dr] & N_{21} \arrow[r] & N_{31} \arrow[r] & N_{41} \\
& N_{22} \arrow[r] \arrow[dr] & N_{32} \arrow[r] & N_{42} \\
& & N_{33} \arrow[r] \arrow[dr] & N_{43} \\
& & & N_{44}
\end{tikzcd} \quad (1)$$

In the operation of the exchange unit, downsampling is achieved using a strided 3×3 convolution, effectively reducing the resolution by selecting a subset of available information. Conversely, upsampling involves a 1×1 convolution followed by nearest neighbor up-sampling. This method increases the resolution while aligning the number of channels, thereby ensuring that the finer resolution maps expand based on the coarser maps, preserving the crucial feature relationships. The inputs to the exchange unit are s response maps, represented as $\{ X_1, X_2, \dots, X_s \}$. Correspondingly, the outputs are also s response maps $\{Y_1, Y_2, \dots, Y_s\}$. Each output, Y_k, is a composite of the input maps, formulated as $Y_k = \sum_{i=1}^{s} a\ (X_i, k)$. Additionally, the exchange unit features an ancillary output map, Y_{s+1}, which is derived as $Y_{s+1} = a\ (Y_s, s+1)$.

B. OmniPose

The OmniPose model [17] is a computer vision and artificial intelligence model constructed using HRNet blocks and the Waterfall Atrous Spatial Pyramid (WASP) v2 module, which is an enhancement of the UniPose model [18]. OmniPose is a significant model in the field of computer vision, designed to address the problem of human pose estimation and skeleton data extraction. It features a flexible and straightforward structure, making it easy to scale and integrate into real-world applications. The OmniPose model utilizes two 3x3 convolutional layers combined with a Resnet-Bottleneck block, resulting in a robust and efficient architecture for processing input data.

Additionally, the model utilizes three blocks of the HRNet model to process diverse data with different resolutions. Each block is accompanied by an enhanced Gaussian heatmap module, aiding in accurately determining the positions of key points on the human body efficiently. Another notable aspect of OmniPose is the assumption that the heatmap of each point follows a Gaussian distribution, optimizing the process of determining the distribution center. This eliminates the need to search for maximum values, thereby significantly improving the model's performance. Furthermore, the model introduces the WASPv2 module, an upgraded version of the WASP module from UniPose, to optimize the classification and localization capabilities of the model. The combination of these components creates a robust and efficient model for human pose estimation and skeleton data extraction. The architecture of the OmniPose model is presented in Fig. 2.

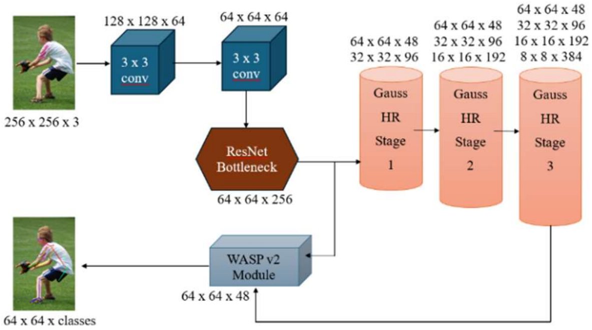

Fig. 2. OmniPose model

In addition to the enhancement method using the SE-Block that we propose in this study, the authors in [21] have also introduced another notable improvement on the OmniPose model, called the Omni-TransPose model. This approach combines the efficiency of the OmniPose model with the effective global information processing capabilities of the self-attention mechanism in the Transformer architecture to improve skeleton data extraction and pose recognition.

Additionally, skeleton extraction models in general and OmniPose in particular can be combined with other powerful object detection models such as ViT, DETR, and YOLO to enhance processing efficiency. A notable example is V-DETR (Pure Transformer for End-To-End Object Detection), a new model proposed by the authors in [22]. V-DETR utilizes the combination of Transformer architectures like DETR and ViT to improve the accuracy and efficiency of object detection. V-DETR can be considered a promising choice for integration with skeleton extraction models to optimize the recognition and processing of image and video data.

C. MPII Dataset

The MPII Human Pose dataset is one of the leading datasets used to evaluate human pose estimation capabilities, particularly in the context of OmniPose model research [17]. This dataset provides a diverse collection of images with a large variety of human activities, serving the purpose of studying and developing accurate and efficient human pose estimation models. This helps clarify how the OmniPose model, enhanced through the integration of attention mechanisms, operates on real-world data and deals with diverse human situations in everyday life.

In addition to providing images of human activities, the dataset also offers detailed information on body part occlusions and 3D orientations of the torso and head. This can assist in evaluating the model's performance in handling complex cases, such as occlusions of body joints or various body angles in space. The diversity and detail of the data in the MPII dataset provide an ideal platform for evaluating and comparing the performance of human pose estimation models, particularly in the context of our proposed research.

The Percentage of Correct Key points (PCK) is one of the commonly used metrics to evaluate the performance of human pose estimation models. This metric was proposed in the research introducing the MPII dataset. The specific formula for calculating the PCK index is as follows:

$$PCK = \frac{TP}{GT} \qquad (2)$$

In the provided definition, TP (true positive) represents the number of key points correctly extracted, while GT (ground truth) is the total number of key points for the object. To assess the accuracy of keypoint localization, the formula r < α * head_bone_length is used, where:

- r is the distance between the extracted key point and the ground truth key point.
- head_bone_length is the distance from the object's head to the extracted key point.
- α is a predetermined coefficient (usually 0.5 or 0.1)

This approach establishes an evaluation standard based on the proximity to the object's head position. When the distance r between the extracted keypoint and the ground truth keypoint is less than α * head_bone_length, the keypoint is considered to be correctly identified. This sets a higher accuracy requirement for key points further away from the object's head position, thereby increasing the flexibility of the evaluation process.

III. SE-BLOCK ARCHITECTURE

The SE-Block was first introduced in the paper Squeeze-and-Excitation Networks [19]. In this study, the authors proposed a new SE (Squeeze-and-Excitation) mechanism to enhance the deep learning capabilities of CNNs by amplifying the importance of significant features and reducing the impact of irrelevant ones. The Squeeze operation aggregates the spatial information of the feature maps to produce a channel descriptor. This is achieved through global average pooling across the spatial dimensions (H × W) of the feature maps (U). The formula for generating the channel-wise statistics z is given by:

$$z_c = F_{sq}(u_C) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_C\,(i,j) \qquad (3)$$

where z_c is the output of the squeeze operation for channel c, and $u_C\,(i,j)$ represents the spatial components of the feature map for that channel

Following the squeeze operation, the excitation operation uses the channel descriptor to produce a set of per-channel modulation weights through a gating mechanism involving two fully connected layers. The excitation formula is defined as:

$$s = F_{ex}(z,W) = \sigma\left(g,(z,W)\right) = \sigma\,(W_2\delta(W_1 z)) \qquad (4)$$

where σ is the sigmoid activation function, δ is the ReLU function, W_1 and W_2) are the weights of the two fully-connected layers involved in creating a bottleneck structure to capture non-linear interactions between channels.

The final output of the SE block is obtained by rescaling the original feature maps U using the weights obtained from the excitation operation. The rescaled feature maps $\tilde{x}_c$ are computed as:

$$\tilde{x}_c = F_{scale}(u_c, s_c) = s_c\, u_c \qquad (5)$$

where s_c is the scalar obtained from the excitation operation for channel c, and u_c is the original input feature map for that channel.

In our research, drawing inspiration from the SE-Block in the aforementioned study, we developed a novel SE-Block to enhance the understanding capability of the OmniPose model in human pose estimation and skeleton extraction, as illustrated in Fig. 3. By integrating the SE-Block into the model, we enhance the attention to crucial features of the image through weighted learning. The integration of the Squeeze-and-Excitation (SE) block into the OmniPose model yields several notable benefits, encompassing aspects of effectiveness, computational performance, scalability, and flexibility, specifically as follows:

• Effectiveness in Feature Enhancement: The SE-Block significantly improves crucial features while reducing irrelevant ones. This feature adjustment enables the model to concentrate more on key points essential for pose estimation, particularly significant in complex pose scenarios.

• Computational Efficiency: One of the primary advantages of the SE-Block lies in its simplicity and efficacy. Unlike conventional attention mechanisms, which may involve complex operations such as dynamic weight adjustment and self-attention, the SE-Block employs a direct method for feature adjustment. This results in lower computational costs, rendering the model more suitable for real-time applications without compromising performance.

• Scalability and Flexibility: The SE-Block seamlessly integrates into various parts of the neural network architecture. This scalability ensures that the enhancements provided by SE-Blocks can be leveraged across different layers and stages of the pose estimation model, contributing to its overall robustness and flexibility.

During the forward process, the SE-Block utilizes a global mechanism to aggregate information from the entire input space into a single vector. Subsequently, this information is passed through a sequence of linear layers to map down to a smaller dimensional space, reducing data dimensions and speeding up computation. Finally, through the Sigmoid layer, the weights are normalized to the range [0, 1], assigning each feature its own importance level. In this way, the SE-Block helps the model focus on important regions in the image and eliminate irrelevant features, thereby improving the understanding and prediction capability of the model in identifying and extracting skeletal data.

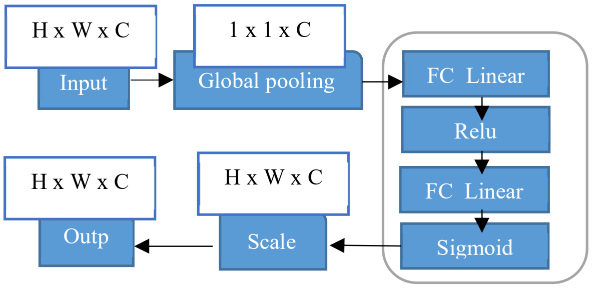

Fig. 3. Structure of SE-Block

Based on the analyzed advantages of SE-Block mentioned above, we have decided to sequentially integrate it into the OmniPose model at the data preprocessing step. Subsequently, experiments were conducted to evaluate the effectiveness of this method. Fig. 4 illustrates the integration of SE-Block into the OmniPose model.

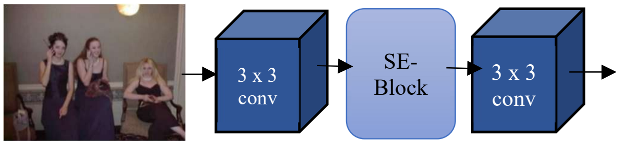

Fig. 4. The integration of SE-Block into the data preprocessing stage in the OmniPose model.

IV. EXPERIMENTS

To comprehensively evaluate the effectiveness of the SE-Block that we proposed when integrated into the OmniPose model for the tasks of human pose recognition and skeletal data extraction, we conducted experiments on the MPII dataset, with 20% for the validation set and 80% for the training set, in two cases: the original OmniPose model and the OmniPose model combined with the SE-Block. The experiments were conducted on the same computer with the CPU configuration Intel I5-13600k, 64GB RAM, and Nvidia 3090 GPU. We used the PCK metric with @0.5 and @0.1 to evaluate and compare. Table I presents the hyper-parameters that we used for training the model in each case.

TABLE I. HYPER PARAMETERS USED FOR EXPERIMENTS PROCESSING

Model	**OmniPose**	**OmniPose with SE-Block**
Training epochs	210	210
Batch size per GPU	16	16
Optimizer	Adam	Adam
Total parameters of model (M)	68.151 M	68.152 M
Momentum	0.9	0.9
Blur-kernel	11	11
LR	0.0001	0.0001
LR-Factor	0.1	0.1
LR-Step	170	170
WD	0.0001	0.0001
Gamma1	0.99	0.99

Table II presents the experimental results which were mentioned above.

TABLE II. THE EXPERIMENTAL RESULTS ON OMNIPOSE WITH SE-BLOCK

Model	**OmniPose**	**OmniPose with SE-Block**
Head	95.430	**96.726**
Shoulder	93.682	**94.684**
Elbow	87.677	**88.308**
Wrist	82.784	**83.041**
Hip	86.394	**87.623**

Model	OmniPose	OmniPose with SE-Block
Knee	83.296	83.216
Ankle	80.066	78.672
PCK@0.5	87.583	**88.028**
PCK@0.1	38.387	37.684

In our research, we evaluated the efficacy of the OmniPose model with SE-Block against the established OmniPose model on the MPII dataset, utilizing the PCK @0.5 and PCK @0.1 metrics for comparison. Previously, the OmniPose model was highlighted as a highly effective approach for human pose estimation, outperforming other models as documented in the original paper [17] and illustrated in Table III and Table IV using the PCK @0.2 metric. Given that our OmniPose with SE-Block model shows enhanced performance in detecting upper body joints compared to the standard OmniPose model, it signifies a notable improvement in the field.

TABLE III. COMPARATIVE PERFORMANCE OF THE OMNIPOSE MODELS: OMNIPOSE (WASPV2), OMNIPOSE (WASP) AND OMNIPOSE (LIGHT)

Model	Omni Pose (WASPv2)	Omni Pose (WASP)	Omni Pose [Light]
Head	**97.40%**	97.40%	96.60%
Shoulder	**97.10%**	96.60%	95.80%
Elbow	**92.40%**	91.90%	89.10%
Wrist	**88.70%**	87.20%	84.30%
Hip	**91.20%**	90.10%	89.00%
Knee	**89.90%**	88.00%	84.10%
Ankle	**85.80%**	83.90%	79.60%
PCK@0.2	**92.30%**	91.20%	89.00%

TABLE IV. THE EXPERIMENTAL RESULTS ON OMNIPOSE AND OTHER MODELS WITH THE MPII DATASET

Model	Omni Pose (WASP v2)	Dark Pose	HRNet	CMU Pose	SPM	RMPE
Head	**97.40%**	97.20%	97.10%	92.40%	92.00%	88.40%
Shoulder	**97.10%**	95.90%	95.90%	90.40%	88.50%	86.50%
Elbow	**92.40%**	91.20%	90.30%	80.90%	78.60%	78.60%
Wrist	**88.70%**	86.70%	86.50%	70.80%	69.40%	70.40%
Hip	**91.20%**	89.70%	89.10%	79.50%	77.70%	74.40%
Knee	**89.90%**	86.70%	87.10%	73.10%	73.80%	73.00%
Ankle	**85.80%**	84.00%	83.30%	66.50%	63.90%	65.80%
PCK @0.2	**92.30%**	90.60%	90.30%	79.10%	77.70%	76.70%

In the calculation formula for $PCK = \frac{TP}{GT}$ presented in section II.C, a key point n is determined as accurate if the distance between the actual joint and the predicted joint is less than the value of α * head_bone_length. From this formula, we can draw the following conclusions:

- For a given value of α, joints that are further from the head of the subject have a larger α * head_bone_length, thus allowing a higher tolerance for discrepancies between the actual and predicted joints. Conversely, joints closer to the head require higher accuracy in recognition. In other words, enhancing the precision in identifying joints in the upper body of the subject is more challenging.
- The α symbol in PCK typically refers to the maximum allowed distance threshold between the estimated and actual joint positions to consider if the joint prediction is accurate. Smaller values of α demand higher accuracy. The value of this threshold can vary depending on the specific needs of the research or application, but common values include: 0.5 – used for initial basic assessments of the effectiveness of the skeletal data extraction method; 0.2 – this threshold balances the requirements for accuracy and recognition capability, suitable for evaluating and fine-tuning the model before deployment; 0.1 – ensures that predictions about the position of key points are extremely accurate, appropriate for advanced research or applications requiring high precision, such as detailed motion analysis.

Based on the observed experimental results, we draw the following conclusions:

- Building on the analysis provided above, enhancing the accuracy of the PCK metric for joints located in the upper part of the body is inherently more challenging. Integrating the SE-Block into the OmniPose model has demonstrated improvements in accurately identifying key point positions near the object's head. However, its effectiveness diminishes for positions further away, such as the knees or ankles. This enhancement is particularly crucial for applications that demand precise recognition of upper body posture.
- The integration of the SE-Block into the OmniPose model significantly enhances the PCK index with α 0.5 compared to the original OmniPose model. However, for the PCK index with α 0.1, the SE-Block does not achieve better performance than the original OmniPose model. The experimental results show that the accuracy rates of the upper body joints are improved, suggesting that the lower PCK scores with α 0.1 originate from poorer recognition of the lower body joints. In other words, integrating the SE-Block into the original OmniPose model enhances the recognition capabilities for upper body joints while diminishing the recognition abilities of lower body joints. This effect becomes more pronounced as the α value in the PCK scale decreases.

V. CONCLUSIONS

In this study, we propose an enhancement method for improving OmniPose model performance in the tasks of skeleton extraction and human pose recognition. This method involves integrating SE-Block into the data preprocessing stage to enhance the importance of critical features. The complexity and training time of the model remains unchanged when applying this solution.

The OmniPose model is chosen for several reasons. Firstly, OmniPose has been proven to achieve high performance in human pose recognition and skeleton data extraction. Additionally, OmniPose utilizes HRNet blocks and the WASPv2 module, providing a flexible and scalable architecture. Notably, OmniPose maintains high resolution throughout the processing, improving the spatial accuracy of key points. Given all these advantages, OmniPose is considered an ideal model for applying various enhancement methods, including the SE-Block. The experimental results on the MPII dataset have demonstrated that integrating SE-Block into the OmniPose model has brought significant improvements in skeleton extraction. This is significant in developing solutions that require high precision in upper-body action recognition. For example, in the realm of sports analytics and coaching for activities that depend heavily on upper-body form-such as swimming, tennis, and table tennis-the ability to accurately assess the movement of joints in the shoulders and arms is crucial for refining techniques and enhancing athletic performance. Furthermore, in the field of medical rehabilitation, particularly for individuals dealing with neck and shoulder injuries or stroke survivors experiencing hemiplegia, the precise monitoring of upper body movements is instrumental. This precision supports the development of tailored rehabilitation practices, thereby facilitating more effective recovery processes for patients.

The enhancement method using the SE-Block is not limited to the OmniPose model but can also be applied to other Human Pose Estimation models. These models typically have convolutional neural network architectures with multiple layers and channels, where channel weight recalibration can provide significant benefits. Models such as HRNet, OpenPose, UniPose, etc. can all benefit from the integration of SE-Block to enhance feature recognition and extraction capabilities. The SE-Block provides a simple yet effective mechanism to improve the ability of CNNs to focus on important features, thereby increasing the accuracy and computational efficiency of the model. HPE models with hierarchical structures and multiple information channels are particularly well-suited for the application of SE-Block, as it helps optimize the learning of crucial features from the input data. Consequently, The solution could be considered when enhancing other models for skeleton extraction and human pose recognition, serving as a basis for improving the effectiveness of solutions in human action recognition; however, its effectiveness may vary depending on the specific model and the context of the problem.

Acknowledgments

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05-2021.04.

References

[1] Johasson G. (1973). *Visual perception of biological motion and a model for its analysis*, Perception and psychophysics, vol. 14, No. 2, pp. 201-211.

[2] M. H. Siddiqi et al. (2021). *A Unified Approach for Patient Activity Recognition in Healthcare Using Depth Camera*, in IEEE Access, vol. 9, pp. 92300-92317, doi: 10.1109/ACCESS.2021.3092403.

[3] Kim, K., Jalal, A. & Mahmood, M (2019). Vision-Based Human Activity Recognition System Using Depth Silhouettes: A Smart Home System for Monitoring the Residents. J. Electr. Eng. Technol, vol. 14, pp. 2567–2573.

[4] Alexander Toshev and Christian Szegedy. *Deeppose: Humanpose estimation via deep neural networks*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1653–1660, 2014.

[5] Alejandro Newell, Kaiyu Yang, and Jia Deng. *Stacked hourglass networks for human pose estimation*. In European Conference on Computer Vision (ECCV), pages 483–499.Springer, 2016.

[6] Shih-En Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. *Convolutional pose machines*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

[7] Bruno Artacho and Andreas Savakis. *Unipose: Unified human pose estimation in single images and videos*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020

[8] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. *Deep high-resolution representation learning for human pose estimation*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019

[9] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. *LCR-Net: Localization-classification-regression for human pose*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017.

[10] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. *MonoCap: Monocular human motion capture using a cnn coupled with a geometric prior*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(4):901–914, 2018

[11] Riza Alp Guler, Natalia Neverova, and Iasonas Kokkinos. *Densepose: Dense human pose estimation in the wild*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7297–7306, 2018

[12] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh. *Realtime multi-person 2D pose estimation using part affinity fields*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[13] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. *Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model*. In European Conference on Computer Vision (ECCV), 2018.

[14] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. *3d human pose estimation with 2D marginal heatmaps*. In IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1477–1485, 2019.

[15] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. *Microsoft COCO: common objects in context*. In ECCV, pages 740–755, 2014.

[16] M. Andriluka, L. Pishchulin, P. V. Gehler, and B. Schiele. *2d human pose estimation: New benchmark and state of the art analysis*. In CVPR, pages 3686–3693, 2014.

[17] Artacho, B.; Savakis, A. *OmniPose: A Multi-Scale Framework for Multi-Person Pose Estimation*. in ArXiv, 2021. https://doi.org/10.48550/arXiv.2103.10180.

[18] B. Artacho and A. Savakis (2020), *UniPose: Unified Human Pose Estimation in Single Images and Videos*, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7033-7042, doi: 10.1109/CVPR42600.2020.00706.

[19] Hu, J., Shen, L., & Sun, G. (2018). *Squeeze-and-excitation networks*. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

[20] Guo, MH., Xu, TX., Liu, JJ. *et al.* Attention mechanisms in computer vision: A survey. *Comp. Visual Media* **8**, 331–368 (2022). https://doi.org/10.1007/s41095-022-0271-y.

[21] Khac Anh – Phu, Van Dung – Hoang, Van – Tuong – Lan Le, Quang – Khai Tran. *Omni-TransPose: Fusion of OmniPose and Transformer Architecture for Improving Action Detection*, 16th Asian Conference on Intelligent Information and Database Systems (ACIIDS 2024).

[22] Dung Nguyen, Van – Dung Hoang, Van – Tuong – Lan Le, *V-DETR: Pure Transformer for End-To-End Object Detection*, 16th Asian Conference on Intelligent Information and Database Systems (ACIIDS 2024).