SAHF-LightPoseResNet: Spatially-Aware Attention-Based Hierarchical Features Enabled Lightweight PoseResNet for 2D Human Pose Estimation

Ali Zakir[1](✉), Sartaj Ahmed Salman[1], and Hiroki Takahashi[1,2]

[1] Graduate School of Informatics and Engineering, Department of Informatics, The University of Electro-Communications, Tokyo, Japan
{a2240012,s2140019}@edu.cc.uec.ac.jp, rocky@inf.uec.ac.jp
[2] Artificial Intelligence Exploration Research Center/Meta-Networking Research Center, The University of Electro-Communications, Tokyo, Japan

Abstract. In recent years, 2D human pose estimation (HPE) has become increasingly important in complex computer vision tasks, including understanding human behavior and interaction. Despite challenges like occlusion, unfavorable lighting, and motion blur, deep learning techniques have revolutionized 2D HPE by allowing automatic feature learning from data and improving generalization. We proposed a new model called Spatially-aware Attention-based Hierarchical Features Enabled Lightweight PoseResNet (SAHF-LightPoseResNet) for 2D HPE. This model extends the simple baseline network by using Spatially-aware Attention-based Hierarchical Features to enhance accuracy while minimizing parameters. The proposed model efficiently captures finer details by incorporating ResNet18, Global Context Blocks, and a novel SAHF module. Our SAHF-LightPoseResNet approach demonstrates superior performance compared to existing state-of-the-art methods, achieving PCKh@0.5 a of 90.8 and a Mean@0.1 metric of 41.1, highlighting its enhanced accuracy and efficiency. This model has important practical applications in robotics, gaming, and human-computer interaction, where accurate and efficient 2D HPE is essential.

Keywords: 2D human pose estimation · SAHF-LightPoseResNet · Global Context Blocks

1 Introduction

The utilization and advancement of Computer Vision (CV) technology in diverse real-world environments, including but not limited to smartphones, digital cameras, and Closed-Circuit Television (CCTV), have led to a persistent flow of extensive data in the form of images and videos. Information regarding human activities found within these data is incredibly significant. Human Pose Estimation (HPE) involves identifying and categorizing the various joints in the human body. Essentially, HPE captures each joint's coordinates, including arms, head, and torso, which are commonly referred to

J. S. Park et al. (Eds.): PDCAT 2023, LNEE 1112, pp. 43–54, 2024.
https://doi.org/10.1007/978-981-99-8211-0_5

as key points that define a person's posture. Over the past few decades, the automatic understanding of HPE has been a major focus of research in CV. 2D HPE offers a fundamental base for several complex CV assignments, including predicting 3D HPE, identifying human actions and motion prediction, parsing human body components, and retargeting human movements. 2D HPE offers extensive support for a wide range of applications, human behaviour understanding, identification of crowd disturbances and riots, detection of violent incidents, recognition of unusual behavior, enhancement of human-computer interaction, and enabling autonomous car driving [4].

The 2D HPE is considered challenging because it is impacted by several significant factors, such as the occlusion of keypoints, unfavorable lighting and background conditions, motion blur, and the complexity of implementing the model in real-world scenarios due to its extensive parameters [20]. Researchers employed conventional techniques like probabilistic graphical models in the early stages to tackle these challenges. However, these methods heavily relied on manually crafted features, restricting the model's generalization ability and limited performance. The progress of 2D HPE has been significantly boosted by introducing deep learning methods, which overcome the generalization limitation of hand-crafted features by enabling automatic feature learning from the data. The remarkable performance of Convolutional Neural Networks (CNNs) in 2D HPE paved the way for developing many deep learning techniques that rely on their success [3].

The main objective of this paper is to achieve high prediction accuracy while minimizing the number of parameters utilized rather than solely focusing on improving the prediction accuracy of existing approaches. The simple baseline network [17] accomplished the best outcomes compared to other top-down approaches. Its effectiveness and simplicity make it an appropriate starting point for creating more sophisticated methods for 2D HPE. In order to accomplish this, we have proposed a unique approach named SAHF-LightPoseResNet, which builds upon the basic framework network by incorporating Spatially-aware Attention-based Hierarchical Features. To reduce the complexity of the model, we have opted to use ResNet18 instead of the more intricate models like ResNet50, 101, or 152, which contain a more significant number of parameters. In our implementation of ResNet18, we have discarded the average pooling segment and the last fully connected segment and exclusively incorporated convolutional layers. Additionally, we have added two deconvolution layers to improve the model's visual processing capabilities and overcome quantization distortion resulting from a large output stride size. We have incorporated Global Context Blocks (GCBs) [2] into the proposed model to equip down-sampler and up-sampler modules with powerful global context features. Our newly developed module, SAHF, merges the feature representations extracted from multiple down-sampler layers, and then enhances these representations by utilizing spatial attention. Finally, SAHF allocates the enriched feature representations to their respective layers in the up-sampler, avoiding conventional skip connections [12, 13]. This approach produced hierarchical representations with spatial awareness and can more effectively capture finer details.

The threefold contribution of the SAHF-LightPoseResNet model can be summarized as follows:

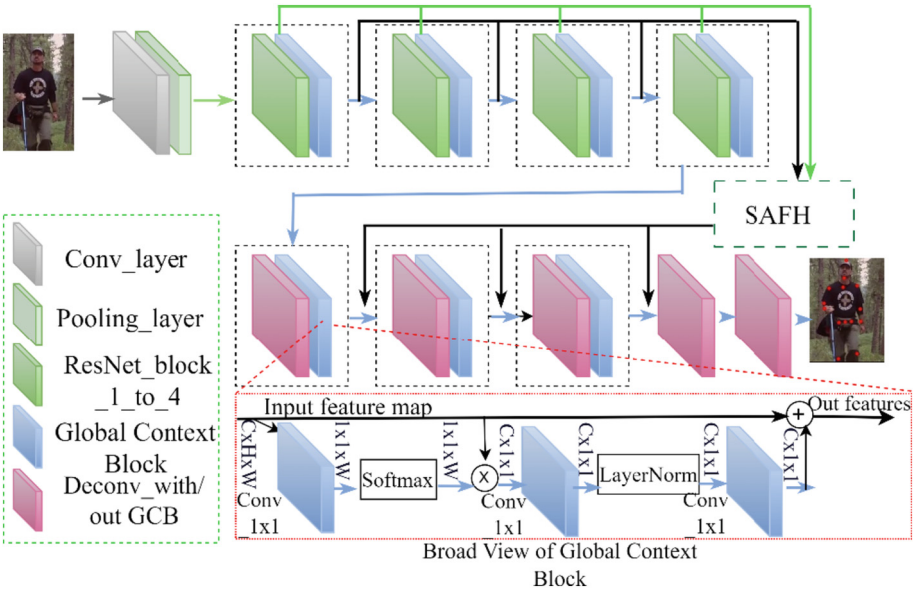


Fig. 1. The proposed SAHF-LightPoseResNet framework.

- We developed a novel model called SAHF-LightPoseResNet, where ResNet18 was used instead of more complex models to reduce the model's complexity. Two deconvolution layers are utilized to improve the model's visual processing capabilities and overcome quantization distortion resulting from a large output stride size. GCBs are incorporated into the model's down-sampler and up-sampler modules to augment it with potent global context features.
- Our proposed SAHF module combines features extracted from different down-sampler layers, which are then enhanced using a spatial attention mechanism and distributed to the corresponding up-sampler layers. As a result, hierarchical representations with spatial awareness are generated, which can capture finer details more effectively.
- Experiments were carried out on the MPII dataset to verify the efficiency of the proposed approach. Evaluation of the quantitative and qualitative outcomes indicated that our model achieved better accuracy and lower computational cost than existing 2D human pose estimation techniques.

This article follows a structured approach with several sections. Section 2 presents an overview of prior research conducted in the same field. Section 3 elaborates on the comprehensive methodology of our Proposed SAHF-LightPoseNet. Section 4 covers pertinent information regarding the experimental setup and implementation details. An analysis of both qualitative and quantitative results is exhibited in Sect. 5. The final section, Sect. 6, draws the conclusion and lays out plans for future exploration.

2 Related Works

Deep learning approaches are utilized in designing network architectures for 2D HPE to extract robust features that span from low to high levels. These approaches are typically categorized into two frameworks: the top-down and bottom-up frameworks. The method of the top-down paradigm involves a sequential process where the initial step is to identify the human bounding boxes in an image, followed by executing the single HPE for every identified box. This type of approach is not a suitable method for managing large crowds as the computational time for the second step increases in association with the number of individuals present [1, 6]. A. Toshev et al. [15] has made a pioneering contribution to the field of HPE by introducing CNN for the first time.

They leveraged the CNN's robust fitting capability to regress the coordinates of human joints and implemented a cascading structure to refine the outcomes continuously. Though, the model tends to overfit because the weights of the fully connected layer depend on the distribution of the training dataset. Convolutional Pose Machine (CPM) [16] and stacked hourglass networks [12] solved this issue by predicting heatmaps of 2D joint locations. Two main object detection techniques exist in 2D HPE: the RCNN [7] series and the SSD series [10]. The RCNN series employs a complicated network structure that achieves high accuracy. Introduced the Mask-RCNN approach, which builds upon the faster-RCNN architecture [7] by incorporating keypoint prediction. As a result, this method achieves excellent results in HPE, demonstrating strong competitiveness in this domain. Conversely, the SSD series offers an average compromise between precision and Y. Chen et al. [5] presents the concept of a cascaded pyramid network (CPN) that uses Global-Net to identify simple keypoints and Refine-Net to handle more challenging keypoints. To be more precise, Refine-Net includes multiple standard convolutional layers that merge feature representations from all levels of GlobalNet.

The process of bottom-up methods start with detecting keypoints for every human instance present in an image. Subsequently, the keypoints of the same individual are joined to form skeletons of multiple instances. This grouping optimization problem is crucial in determining the outcome of the bottom-up approach. Some representative methods utilize this approach, and they are [3, 14]. Open-Pose, as described in [3], utilized two branches - one of which employed a CNN to predict all keypoints based on heatmaps, and the other used a CNN to acquire part affinity fields. The part affinity fields represent 2D direction vectors, and they serve as a confidence metric to determine if the keypoints are associated with the same person. Ultimately, both branches are merged to generate the concluding prediction. The approach known as associative embedding [11], derived from Hourglass [12], is end-to-end trainable. The source detected and accumulated keypoints in one step without requiring two separate processes. Implementing bottom-up approaches can be challenging due to the difficulty of combining information from multiple scales and grouping features together. Even with the introduction of effective grouping procedures, these methods still struggle to contest top-down strategies for pose estimation. In recent times, the majority of cutting-edge outcomes have been achieved through top-down methodologies. Our research traced the top-down approach and developed a successful 2D HPE model. This addresses the issue of top-down approaches by modifying a baseline network with Spatially-aware Attention-based Hierarchical Features. We utilized a simpler ResNet18 model and removed specific layers

to reduce complexity. We then added deconvolution layers and Global Context Blocks to improve visual processing and global context features. The proposed SAHF module combines and enhances feature representations from various layers, enabling better capture of finer details through hierarchical representations with spatial awareness.

3 Our Proposed SAHF-LightPoseNet

To formally define the task of estimating human pose, we can state it as follows: when given an RGB image or video frame I as input, the goal is to estimate pose P of human(s) present in the data. The pose P can be represented as a set of K's keypoint positions, where a two-dimensional coordinate represents each keypoint (x_k, y_k), and K can vary depending on the dataset. Therefore, we aim to estimate the pose $P = \{P_i\}_{i=1}^{n}$ for all n individuals in the input data. Our research builds upon the simple baseline network for 2D HPE that was previously developed. Our proposed SAHF-LightPoseResNet is shown in Fig. 2. Further information and comprehensive explanations regarding the components of SAHF-LightPoseResNet are introduced in the following subsections.

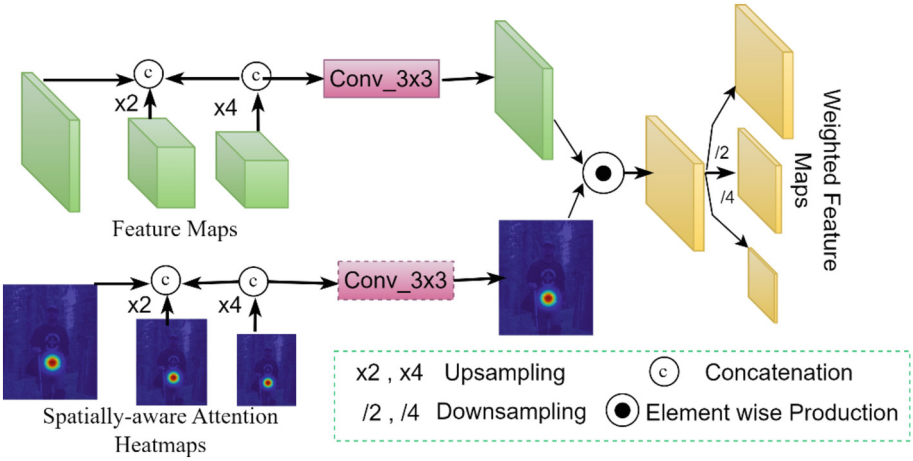


Fig. 2. Proposed SAHF Module.

3.1 Enhancing Backbone Model with Modified ResNet and Deconvolution Module

The structure of the autoencoder network is commonly utilized for dense labeling tasks. To achieve this, we employed an autoencoder network structure that slowly decreases the resolution of embeddings to capture extended-range details, which subsequently increases feature maps while recovering spatial resolution. Hourglass and simple baseline networks create smaller output feature maps than their input feature maps, which are then resized using a simple transformation technique that can cause quantization errors.

When data processing is biased, prediction errors can occur due to horizontal flipping and how the model processes the output resolution [9].

We incorporated two deconvolution modules into our approach to tackling the above-mentioned challenges. These modules were designed to generate a complete output feature map and were integrated within the architecture of the simple baseline network. We opted to use ResNet 18 and 34, which have fewer parameters compared to more complex ResNet models like 50, 101, or 152. We modified ResNet [8] by removing the average pooling segment and fully connected part and replacing them with four ResNet blocks after a convolutional and pooling layer. The first set of layers in the network, which includes a convolutional layer and a pooling layer, reduces the size of the feature maps by half. As the input passes through each block of the network, additional convolutional layers are used to decrease the feature maps by two strides while simultaneously increasing the number of filters by a factor of two. We added five deconvolutional modules with batch normalization and Hardswish activation, each doubling the feature resolution map until the output matches the input. The proposed architecture is illustrated in Fig. 1. The 4th and 5th deconvolutional layers have channel sizes of 128 and 64, respectively.

3.2 Amplifying Model Performance with Global Context Blocks

In computer vision, a global context block is a module designed to capture the overall spatial information of an input feature map, aiming to improve object recognition in an image. In convolutional layers, the association among pixels is only considered within a local neighborhood, and baseline network. We opted to use ResNet 18 and 34, which have fewer parameters compared to more complex ResNet capturing long-range dependencies requires multiple convolution layers. To address this limitation, researchers proposed a non-local operation [18], which employed a self-attention mechanism from [19] to model long-range dependencies. Using a global network creates an attention map tailored to each query position, enabling the collection of contextual features that can then be integrated into the features of the corresponding position. GCNet is presented as a highly well-organized and operative method for global context modeling [2]. This method employs a query-agnostic attention map to generate a contextual representation that can be globally shared and then incorporates it into the features of each query location in the network.

Our proposed method uses global context blocks [2] to enhance the spatial information of input feature maps. Specifically, as illustrated by blue blocks in Fig. 1 global context blocks are incorporated into each ResNet block as well as the first three blocks of the deconvolution modules. We generate a spatially-aware attention heatmap using a 1×1 convolution and SoftMax to produce attention weights, which are then used in attention pooling to extract a global context feature. Channel-wise dependencies are obtained using the bottleneck transform technique. Afterward, the resulting global context features are combined with the features of each position in the network, as shown in the following equation.

$$f_g = \sum_{i=1}^{h} \sum_{j=1}^{w} w_{ij} f_{ij} \tag{1}$$

where in Eq. 1, f_g represent the global context feature, *h and w* are the height and width of the input feature map, w_{ij} is the attention weights at position (i, j) and $f(i, j)$ is the feature vector at the position (i, j).

3.3 SAHF Module

The Spatially-aware Attention-based Hierarchical Features (SAHF) module overcomes the limitations of earlier frameworks, such as the simple baseline framework, which did not integrate skip connections [12, 13]. These connections have proven effective in U-Net and hourglass networks for retaining spatial information at each feature map, allowing for an efficient transfer of spatial information across the network, leading to improved localization.

Our proposed SAHF module, depicted in Fig. 2, is an alternative to traditional skip connections used in previous works [12, 13]. The SAHF module combines hierarchical features from different layers, using spatial attention to enhance features. It receives feature maps from the first three Global Context Blocks, ResNet blocks, and Spatially-aware attention feature maps. These feature maps are multiplied elementwise to generate enhanced features, which are then allocated to the deconvolution modules, excluding the last one. The Spatially-aware attention technique focuses on locations related to pose estimation and helps generate helpful detail while suppressing background information. The enhanced features from the SAHF module improve the capabilities of related deconvolution models, leading to an overall improvement in network performance as shown in table 1 and visualize the performance of SAHF in Fig. 3 (a) and (b).

3.4 Heatmap Joint Prediction

Our model predicts joint positions at the pixel level by converting them into heatmaps within a bounding box, using a 2D Gaussian function to generate ground truth. The resulting heatmap represents the probability of a joint being located at each pixel.

$$H_{k(x,y)} = exp(\frac{-[(x - y_k)^2 + (y - y_k)^2]}{2\sigma^2} \tag{2}$$

In Eq. 2 H_k represent heatmap for kth joint where $k \in \{1, 2, \ldots, K\}$, and (x, y) show the position of the specified pixel in the heatmap. The k^{th} joints coordinated are denoted by (x_k, y_k). The value for spatial variance σ is set to 12 in this experiment.

4 Experimental Setup

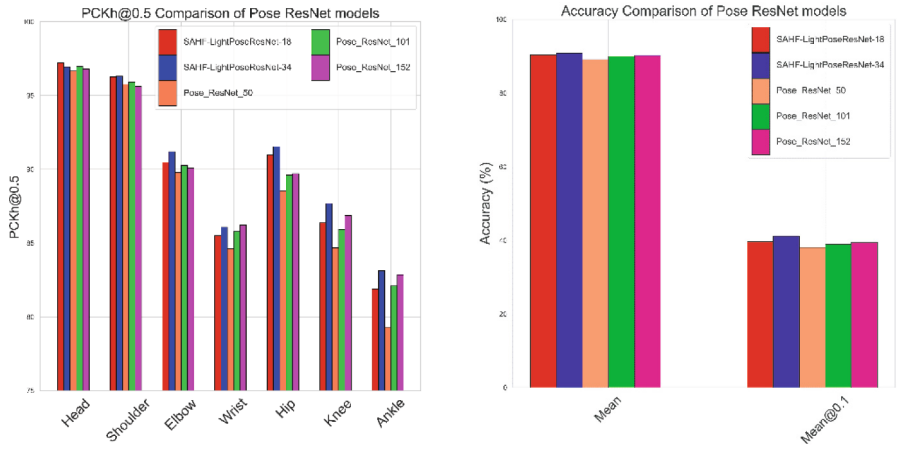
4.1 Dataset

Our experimentation to evaluate the effectiveness of our proposed model was carried out using the extensively recognized MPII (Max Planck Institute for Informatics) Dataset [21]. This comprehensive dataset comprises more than 25,000 annotated images, capturing over 40,000 individuals, each mapped with 16 distinct key-points. This substantial data set has been strategically split into two subsets, one for training and the other for

testing. A total of 28,000 images were employed to build and refine our model in the training phase. Subsequently, a separate set of 11,000 images was exclusively leveraged to test the model's performance, providing an objective measure of model's robustness and accuracy.

4.2 Implementation Details

We utilized data augmentation techniques to improve the model's ability to handle scale variance and spatial rotation, including random horizontal flip, rotation within -40 to + 40 degrees, and scaling between 0.7 to 1.3 in our approach. Our designed model was implemented using PyTorch. The training process included a learning rate of 1e-05, a batch size of 16, a number of workers set to 6, and 150 epochs.



(a) **(b)**

Fig. 3. (a) Illustration of PCKh@0.5 Results: Proposed Model and Simple Baseline Models (b) Graphical Analysis of Mean and Mean@0.1: Proposed Models and Simple Baseline

In our research, we implemented three key components to ensure precise model training and enhanced model performance: MSE loss function, AdamW optimizer, and Hard Swish activation function.

The Mean Square Error (MSE) loss function, which has been effectively utilized in previous works such as [6, 17], was chosen to evaluate the model's error. The formula for MSE is presented below.

$$L = \frac{\sum_{k=1}^{K} \|H_k - \widehat{H}_k\|}{K} \tag{3}$$

In Eq. 3, $\widehat{H}_k$ represent estimated heatmap for the k$^{\text{th}}$ joint, where as H_K is the heatmap for the k$^{\text{th}}$ joint $k \in \{1, 2, \ldots, K\}$.

The model optimization process was further enhanced using a variant of the Adam optimizer, AdamW. Distinct from the original Adam optimizer, AdamW separates weight decay from the learning rate, enabling independent optimization and significant reduction in overfitting.

Finally, our research employed the Hard Swish activation function. This superior function offers significant advantages over the commonly used ReLU function, including superior accuracy, efficiency, smoother gradient, and the ability to address the 'dying neurons' issue often seen with ReLU. Utilizing Hard Swish, we witnessed an overall performance improvement in our neural network model and achieved superior experimental results, suggesting its potential benefits across various deep-learning applications.

4.3 Evaluation Metrics

We used PCK (Percentage of Correct Keypoints) and Mean@0.1, widely used evaluation metrics in HPE tasks. PCKh, a variation of PCK, compares the predicted and actual keypoints using the head bone link length as a reference. The prediction is considered correct if the distance between the predicted and actual keypoints is less than 50% of the head bone link length (PCKh@0.5). Mean@0.1, on the other hand, measures the average distance between predicted and actual keypoints, normalized by the head bone link length, and is scale-invariant.

5 Experimental Results and Discussion

Our model was trained on different input sizes, namely 256 × 256, 288 × 384, and 384 × 384. The only exception was the simple baseline model, which did not use the 288 × 384 input size. In Fig. 4, you can observe the inference results of the LightPoseResNet-18 model on the MPII dataset. To compare the performance of our proposed SAHF-LightPoseResNet model with that of the basic baseline model, we present their outcomes in Table 1. We also provide a visualization of each joint with PCKh@0.5 for our proposed models and the simple baseline models that used the 256 × 256 input size, as shown in Fig. 3(a). Finally, Fig. 3(b) displays both models' overall Mean and Mean@0.1 predictions using the 256 × 256 input size. Initially, we conducted training on SAHF-LightPoseResNet-18 using input sizes of 256 × 256. As a result, we are able to obtain PCKh@0.5 values of 89.425 and 90.297, along with mean@0.5 values of 34.483 and 39.670. These values were found to be higher than the PCKh@0.5 and Mean@0.1 values of all the basic baseline models. We conducted an experiment on the LightPoseResNet-18 model using input sizes 288x384 and 384 × 384. Our results showed that the LightPoseResNet-18 model outperformed the simple baseline. Notably, despite achieving better results, the LightPoseResNet-18 model used only 21 million parameters during the training process, which is fewer than all the simple baseline models.

Some experiments were conducted on LightPoseResNet-34 using the input sizes mentioned earlier. The outcomes revealed that the model yielded better results in terms of PCKh@0.5 and Mean@0.1, despite having fewer parameters than the simple baseline model. These findings are presented in Table 1 and visualized in Fig. 3(a) and Fig. 3(b).

Fig. 4. Qualitative Results on MPII pose estimation result, containing viewpoint change, and occlusion and self-occlusion.

Therefore, our study demonstrates the effectiveness of the LightPoseResNet-18 and 34 models in terms of both computation and performance.

Table 1. Performance comparisons of our SAHF-LightPoseResNet with simple baseline results on MPII dataset

Model	No.par	input	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean	Mean@0.1
Pose_Resnet_50	34.0M	256x256	96.351	95.329	88.989	83.176	88.420	83.960	79.594	88.532	33.911
		384x384	96.658	95.754	89.790	84.614	88.523	84.666	79.287	89.066	38.046
Pose_Resnet_101	53.0M	256x256	96.862	95.873	89.518	84.376	88.437	84.486	80.703	89.131	34.020
		384x384	96.965	95.907	90.268	85.780	89.597	85.935	82.098	90.003	38.860
Pose_Resnet_152	68.6M	256x256	97.033	95.941	90.046	84.976	89.164	85.311	81.271	89.620	35.025
		384x384	96.794	95.618	90.080	**86.225**	89.700	86.862	82.853	90.200	39.433
SAHF-LightPoseResNet_18	21.0M	256x256	96.965	95.688	89.398	84.051	90.254	85.029	80.728	89.425	34.483
		288x384	97.169	95.788	90.131	84.462	90.341	85.331	81.696	89.766	36.435
		384x384	97.203	96.264	90.472	85.489	90.981	86.379	81.890	90.297	39.670
SAHF-LightPoseResNet_34	30.0M	256x256	97.237	95.805	90.012	84.891	90.064	85.976	81.507	89.846	36.417
		288x384	**97.271**	96.247	90.608	85.642	91.016	86.984	82.712	90.536	38.158
		384x384	96.930	**96.298**	**91.188**	86.072	**91.535**	**87.668**	**83.137**	**90.877**	**41.137**

6 Conclusion and Future Work

In this research work, we proposed SAHF-LightPoseResNet for 2D HPE. The SAHF-LightPoseResNet model is a novel approach utilizing ResNet18 to reduce complexity while achieving effective visual processing capabilities. The model's down-sampler and upsampler modules are enhanced with GCBs to provide potent global context features. The SAHF module combines and distributes features with spatial attention to produce hierarchical representations with spatial awareness that capture finer details effectively.

SAHF-LightPoseResNet performs better than basic baseline models on the MPII dataset due to improved features, better activation function, and advanced model optimizer, as simulation results indicate. In the future, our model can be utilized for 3D human pose estimation with object recognition and hand pose estimation due to its general applicability.

References

1. Bertasius, G., Feichtenhofer, C., Tran, D., Shi, J., Torresani, L.: Learning temporal pose estimation from sparsely-labeled videos. In: Advances in Neural Information Processing Systems 32 (2019)
2. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: Gcnet: non-local networks meet squeeze excitation networks and beyond. In: Proceedings of the IEEE/CVF international conference on computer vision workshops, p. 0 (2019)
3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299 (2017)
4. Chen, H., Feng, R., Wu, S., Xu, H., Zhou, F., Liu, Z.: 2D human pose estimation: a survey. Multimedia Systems, pp. 1–24 (2022)
5. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7103–7112 (2018)
6. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2334–2343 (2017)
7. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
9. Huang, J., Zhu, Z., Guo, F., Huang, G.: 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, pp. 5700–5709 (2020)
10. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
11. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: Advances in Neural Information Processing Systems 30 (2017)
12. Alejandro Newell, Kaiyu Yang, Jia Deng,: Stacked hourglass networks for human pose estimation. In: Bastian Leibe, Jiri Matas, Nicu Sebe, Max Welling, (ed.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
13. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Salman, S.A., Zakir, A., Takahashi, H.: Cascaded deep graphical convolutional neural network for 2D hand pose estimation. In: International Workshop on Advanced Imaging Technology (IWAIT) 2023. vol. 12592, pp. 227–232. SPIE (2023)

15. Toshev, A., Szegedy, C.: Deeppose: human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1653–1660 (2014)
16. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4724–4732 (2016)
17. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European conference on computer vision (ECCV), pp. 466–481 (2018)
18. Wang, X., Ross, G., Abhinav, G., He, K.: non local neural networks. In: Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition, pp. 7794–7803. (2018)
19. Vaswani, A., et al.:Attention is all you need. In: Advances in Neural Information Processing Systems 30 (2017)
20. Zheng, C., et al.: Deep learning-based human pose estimation: a survey. arXiv preprint arXiv: 2012.13392 (2020)
21. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: In: 2D human pose estimation: new benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, pp. 3686–3693 (2014)