IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 August 2024, accepted 30 September 2024, date of publication 8 October 2024, date of current version 21 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3476196

RESEARCH ARTICLE

WideHRNet: An Efficient Model for Human Pose Estimation Using Wide Channels in Lightweight High-Resolution Network

ESRAA SAMKARI, MUHAMMAD ARIF, MANAL ALGHAMDI, AND MOHAMMED A. AL GHAMDI, (Member, IEEE)
Department of Computer Science and Artificial Intelligence, Umm Al-Qura University, Makkah 21955, Saudi Arabia
Corresponding author: Muhammad Arif (mahamid@uqu.edu.sa)

The authors extend their appreciation to the Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number: IFP22UQU4250002DSR226.

ABSTRACT Human pose estimation is a task that involves locating the body joints in an image. Current deep learning models accurately estimate the locations of these joints. However, they struggle with smaller joints, such as the wrist and ankle, leading to lower accuracy. To address this problem, current models add more layers and make the model deeper to achieve higher accuracy. However, this solution adds complexity to the model. Therefore, we present an efficient network that can estimate small joints by capturing more features by increasing the network's channels. Our network structure follows multiple stages and multiple branches while maintaining high-resolution output along the network. Hence, we called this network Wide High-Resolution Network (WideHRNet). WideHRNet provides several advantages. First, it runs in parallel and provides a high-resolution output. Second, unlike heavyweight networks, WideHRNet obtains superior results using a few layers. Third, the complexity of WideHRNet can be controlled by adjusting the hyperparameter of expansion channels. Fourth, the performance of WideHRNet is further enhanced by adding the attention mechanism. Experimental results on the MPII dataset show that the WideHRNet outperforms state-of-the-art efficient models, achieving 88.47% with the attention block.

INDEX TERMS Convolution neural network, efficient network, human pose estimation, wide network.

I. INTRODUCTION

Many computer vision tasks use deep learning to boost performance [1], [2], and the human pose estimation task is no exception. Human pose estimation involves identifying the positions of body joints (e.g., head, neck, wrist, elbow) in the image. The network model used for this task estimates the coordinates (x, y) for each keypoint (joint) in the input. This task has numerous applications, including abnormal behavior detection [3], [4], pose tracking [5], [6], and gesture translation [7]. In addition, it can serve as a fundamental step for other tasks such as 3D human reconstruction [8].

Among deep learning types, conventional neural networks (CNN or ConvNet) work well with the image [1], and therefore, most studies of human pose estimation use CNN [2]. Unlike the traditional method that requires handcraft features, CNN helps extract robust features without needing an expert engineer to do feature engineering. Hence, CNN shifts the researcher's work to focus more on building the network structure rather than doing handcraft features. Fig. 1 shows some of these handcraft features.

One approach to creating a strong ConvNet structure is to increase the number of layers, making the network deeper. Deeper ConvNets can capture more complex features [9], [10], but they also face issues such as exploding or vanishing gradients during training. One solution to this problem is the use of skip connections [11]. Another challenge with deep networks is their increased complexity and high computational demands [12], [13], [14], [15], [16], which makes them unsuitable for lightweight devices.

The associate editor coordinating the review of this manuscript and approving it for publication was Domenico Rosaci.

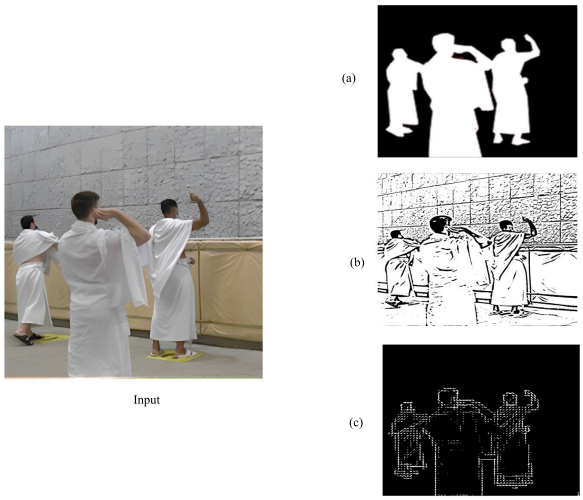


FIGURE 1. Examples of handcrafted features, where (a) silhouette, (b) edges, and (c) histogram of oriented gradients.

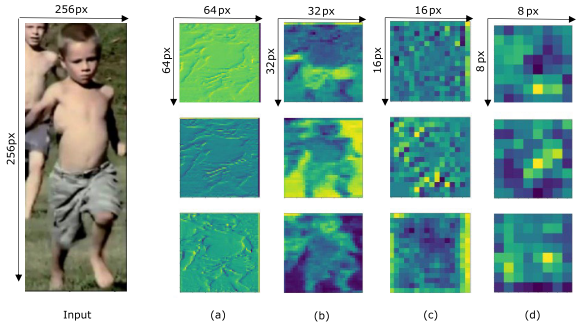


FIGURE 2. Multiscale resolutions of feature maps. Models of human pose estimation utilize different feature resolutions to enrich the information of the target features. Here, we illustrate four different resolutions of feature maps, where (a) is high resolution, (b) and (c) are medium resolution, and (d) is low resolution. These feature maps are extracted from the WideHRNet.

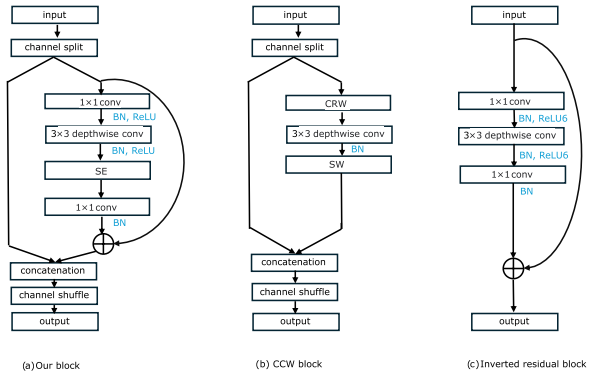


FIGURE 3. Building block, where (a) is the proposed block that is inspired by various blocks, including (b) the conditional channel weighting (CCW) and (c) the inverted residual block. The stride value of all these blocks is 1. Conv: convolution, BN: batch normalization, SE: squeeze-excitation block, CRW: cross-resolution weights block, and SW: spatial weights block.

To address these issues, using a wide network can be more effective. A wide network has fewer layers than a deep network but includes more neurons (channels) in each hidden layer [11], [17]. This design allows the model to operate efficiently on devices with limited resources [11], [17], [18].

According to Cheng et al. [19], the major difference between the deep and wide networks is that deep networks excel at generalization, while wide networks are better at memorization. Although memorization can lead to issues like overfitting [10], wide networks have demonstrated superior performance in various tasks, including classification [9], [20], detection [11], [17], and single-image super-resolution [10].

Inspired by these wide networks, this paper aims to build a wide network for pose estimation by increasing the number of channels in the hidden layer. However, the models of human pose estimation require fusing multiresolution features to obtain a high accuracy [12], [16], [21] and existing wide networks [9], [10], [11] did not have entirely focusing on this point. In addition, many literature reviews [1], [2] show that the networks that combine multiresolution outperform the other pose estimation networks. For example, models that use a multiresolution such as Stack Hourglass [16] and High-Resolution Network (HRNet) [12] are widely used over other networks like the Residual Network (ResNet) [1], [2]. Fig. 2 illustrates what multiresolution looks like.

Recently, HRNet has been widely adopted by numerous studies focused on pose estimation [14], [15], [22], [23]. The HRNet architecture consists of multiple stages and branches, which contribute to its high accuracy. Additionally, HRNet improves performance by adding more layers, creating a deeper model. However, this makes it less suitable for lightweight devices. With the growing interest in developing small and efficient networks [11], [17], [24], [25], [26], models like Small HRNet [23] and Lightweight HRNet (LiteHRNet) [22] have been proposed. These models enhance HRNet's efficiency by reducing the network's depth and width. Among these networks that fuse multiscale features, we adopted the LiteHRNet model and modified it to make it wider.

To make LiteHRNet wider, we first need to understand its basic block. LiteHRNet is state-of-the-art in building an efficient model for the human pose estimation task. It follows the HRNet structure in terms of having multiple stages and branches. However, it replaces the residual block with an efficient CNN block called Conditional Channel Weighting (CCW). The CCW block contains several efficient components, as shown in Fig. 3(b). However, based on our analysis, we observed that only the depthwise convolution in CCW has a significant impact on the accuracy. Therefore, we removed all blocks except the depthwise layer in the CCW. Then, we increased the number of channels before applying the depthwise convolution. By increasing the channels, we made the network wider. Thus, we named the new model WideHRNet.

As shown in Fig. 3(a), the basic block of WideHRNet increases the number of channels by using a pointwise layer

(1×1 convolution) before performing any convolution. This expanding method follows the inverted residual block [11] in terms of controlling the expansion ratio and building a linear bottleneck; see Fig. 3(c). In addition, we utilize the Squeeze-Excitation (SE) block [27], the attention mechanism technique, to extract more important features between the expanding channels. Furthermore, light operations such as channel split, skip connection, and channel shuffle are also used.

We can summarize the contribution of the paper as follows:

- We leverage parallel performance in building a wide and efficient model that requires less amount of complexity. Hence, we present a WideHRNet for human pose estimation. In each layer in WideHRNet, the model increases the number of channels to increase the ability to capture more features of small joints.
- We use different techniques to enhance the performance of WideHRNet. In this paper, we use a channel split, channel shuffle, and shortcut connection to reduce complexity, improve accuracy, and avoid the training problem, respectively. All of these techniques did not add complexity.
- We add the attention mechanism to the expanding channels block. Specifically, we use the Squeeze-Excitation attention block to weight the expanding channels. The experiment shows that adding channel attention has enhanced accuracy without adding complexity to the model.
- Our experiment on the MPII dataset shows that the WideHRNet model performs better than the LiteHRNet model and other efficient models.

The rest of the paper is organized as follows: Section II provides an overview of state-of-the-art networks used in human pose estimation. Next, Section III explains in detail the structure and basic block of the WideHRNet model. Section IV presents the results of the WideHRNet and discusses the effect of the increase in the number of channels on performance. Section V presents a conclusion with limitations and future works of this paper.

II. RELATED WORK

Many works concern estimating human joints using ConvNet (i.e. [28], [29], [30], [31], [32]). Building a deep network is one way to achieve high accuracy for the pose estimation task [12], [16]. However, this method is not suitable for lightweight devices. Therefore, recently, many studies [11], [17], [18] have focused on building an efficient network by making the network wider rather than deeper. This section provides an overview of the state-of-the-art models and divides the topics into two subsections: heavyweight networks and lightweight networks. In addition, it covers the attention mechanism techniques.

A. HEAVYWEIGHT NETWORKS

Many studies [12], [15], [33] focus on developing high-accuracy models using different techniques. For instance, the Stacked Hourglass model [16] uses a repetitive symmetric low-to-high resolution with adding a skip connection to preserve spatial information. This model has been adopted by many studies [34], [35]. Later, however, Sun et al. [12] noted that recovering a high-resolution representation from a low-resolution representation will likely affect the quality of the predicted output. Hence, they proposed HRNet, which has a parallel architecture. In the HRNet model, the high-resolution representation is preserved along the network while gradually adding high-to-low-resolution sub-networks in parallel. HRNet has been adopted by many studies [13], [14], [36], [37], and one such study is HRFormer [15] that integrates the Swin Transformer with CNN. Both Hourglass and HRNet achieve high accuracy by making the network deeper, i.e., by increasing the number of hidden layers.

Unlike CNN, which is concerned with capturing local-range dependence, the Transformer helps capture long-range dependence [38], [39]. Therefore, recent studies [33], [40] have developed models that rely entirely on Transformer to estimate human pose, which is currently considered one of the most accurate models for human pose estimation.

Despite the achievements of all the above-mentioned models, they are unsuitable for lightweight devices due to their high computation complexity.

B. LIGHTWEIGHT NETWORKS

Small networks allow for estimating the human pose on less powerful devices. Therefore, there are many studies [11], [17], [22], [23], [25], [26] focused on designing an efficient networks. To build lightweight networks, many methods that provide a low computational are used, such as dynamic CNN [41], [42]. Furthermore, some works [24], [25], [26] use techniques such as attention mechanism and channel shuffle to boost the performance. Another popular method is to use an efficient CNN block with fewer layers and more channels. For instance, MobileNetV1 [17] and MobileNetV2 [11] use a depthwise separable convolution (depthwise conv followed by a 1×1 conv) with network architecture search (NAS) to build an efficient and wide network.

Unlike the previous models [11], [17], [25], [26] which were primarily developed for classification tasks and later some of them were used for pose estimation [11], [24], [26]. The LiteHRNet [22] model was specifically created for human pose estimation. The LiteHRNet model achieved good results by utilizing the Small HRNet [23] as the backbone and replacing its basic block with ShuffleV2 block [26]. However, the ShuffleV2 block adds complexity to the LiteHRNet; hence, all the complex convolution layers in the ShuffleV2 block were replaced with less complex convolution. Therefore, the LiteHRNet has become a state-of-the-art model in pose estimation task [41], [43], [44].

Later, many models have improved LiteHRNet by using a dynamic CNN block [41], [42] or adding different attention mechanisms [43], [45]. However, they either made their architecture difficult to follow despite its performance [41],

[42] or made minor improvements in accuracy [43], [45]. Therefore, this paper aims to propose a block with a simple structure, high accuracy, and low complexity. This is accomplished by extending LiteHRNet, enhancing its width, and incorporating an attention mechanism to improve performance further.

C. ATTENTION MECHANISM

The attention mechanism helps to improve the performance without affecting the complexity [27], [46]. There are six categories of attention [47], and among them is channel attention. Channel attention [47] can be defined as "what to pay attention to". It helps to select important channels.

The first study that suggest the channel attention was Squeeze-and-Excitation Network (SENet) [27]. The main idea is to compress all channels as a single value and then use these values to reweight the channels to distinguish between the channels and understand which one of the channels has the most important information. Following this work, several studies [46] have proposed other channel attentions, for example, Effective Channel Attention (ECA) proposes to replace dimensionality reduction in SENet with a local channel interaction strategy.

Since the wide model means increasing the number of channels, one of the channel attention techniques should be used. Therefore, in this paper, we use the SE channel attention to weight each channel in the proposed block.

III. METHODS

This section describes the proposed model. First, we will describe the structure of the LiteHRNet model that we modified and proposed our model based on. Then, we will explain what layers we used to build our block and why this block structure follows a linear bottleneck. Next, we will discuss the channel attention block used in the proposed block. Finally, we will describe the structure of the WideHRNet model. Fig. 4 shows the general structure of WideHRNet.

A. LITEHRNET MODEL

We chose LiteHRNet because it is an efficient model that delivers good performance with low complexity for the pose estimation task. LiteHRNet structure follows the HRNet model, a design that includes multiple stages and branches. In each stage, information from different branches is exchanged through a multiscale fusion unit, which helps the high-resolution subnetwork receive more information. Hence, the predicted heatmap (output) is more likely to be high quality, resulting in more accurate detection of the keypoints.

To further enhance performance, the LiteHRNet follows the Small HRNet in terms of using fewer layers with a smaller width. In addition, the LiteHRNet model introduced a Conditional Channel Weighting (CCW) block to improve the performance; see Fig. 3(b).

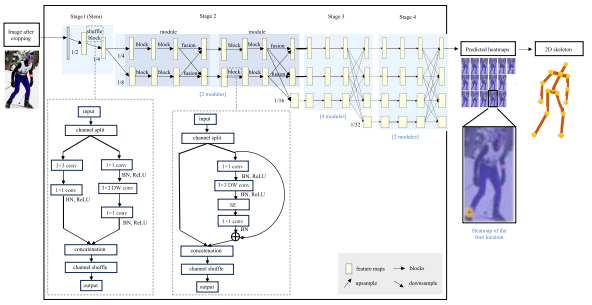


FIGURE 4. WideHRNet model structure. We adopted the LiteHRNet model and modified its main block, replacing its sub-blocks with a residual block and channel attention. This new block is used in all stages except stage 1, which uses the ShuffleV2 block. In this figure, we only visualize the number of blocks and modules for stage 2. For more information about the other stages, we refer to Table. 2. Conv: convolution, BN: batch normalization, DW: depthwise, and SE: squeeze-excitation block.

TABLE 1. Ablation study on conditional channel weighting (CCW) of LiteHRNet_18 model. The MPII val set is used for evaluation, where the input size is 256 × 256. CRW: cross-resolution weights block, DW: depthwise layer, and SW: spatial weights block.

Model	CRW	DW	SW	#Params (M)	FLOPs (G)	PCKh
w/ CCW block	✓	✓	✓	1.1	0.27	86.1
w/o CCW block	✗	✗	✗	0.90	0.22	82.0
Multiresolution weights	✓	✗	✗	0.99	0.24	83.3
3 × 3 depthwise convolution	✗	✓	✗	0.90	0.25	85.3
Spatial weights	✗	✗	✓	0.96	0.23	83.4
CCW w/o SW	✓	✓	✗	1.02	0.27	85.9
CCW w/o CRW	✗	✓	✓	0.99	0.25	85.5
CCW w/o DW	✓	✗	✓	1.1	0.25	83.1

The CCW block splits the input channels into two branches: one branch is used as identity, whereas the other branch contains three sub-blocks: cross-resolution weights (CRW), 3 × 3 depthwise convolution, and spatial weights (SW). CRW block applies element-wise weighting operation between the input channels and channels from different resolutions. Hence, it weights the maps by exchanging information from all the input channels of all the resolutions. Then, a 3 × 3 depthwise convolution is performed to do some filtering. Finally, the SW block is used to reweight the feature map channels.

Table. 1 shows how each of these sub-blocks affects the performance of the model. According to Table. 1, the depthwise convolution layer has a significant effect compared to other blocks. Hence, we attempt to utilize this layer and boost the performance by making the depthwise layer receive more channels.

B. LINEAR BOTTLENECK

Adding an activation function before any convolution layer may result in some information loss [11]. The activation functions set some threshold to perform the nonlinear operation. For example, the ReLU activation sets a threshold that changes all negative inputs to zero. This allows the model to learn nonlinear mappings between inputs and outputs. However, setting a certain threshold may cause some information to be lost. Therefore, we address this problem

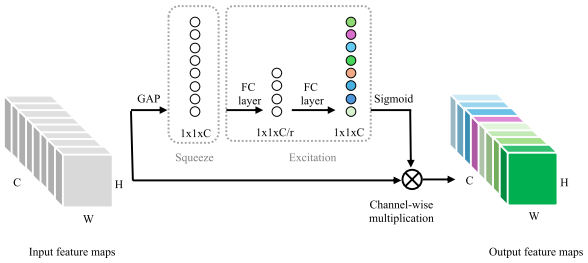


FIGURE 5. Squeeze-Excitation block. It consists of two modules: squeeze and excitation. The squeeze module contains a global average pooling (GAP) layer, whereas the excitation module contains two fully connected (FC) layers. This block is used to weight the input channels. C: channel, W: width, and H: height, r: reduction ratio.

by increasing the number of channels before applying any nonlinear operations. Hence, if information is lost, it can be recovered via other channels.

Our block consists of three convolutional layers: two pointwise and one depthwise. First, a pointwise convolution is applied to increase the number of channels. Then, batch normalization is performed, followed by a nonlinear activation function. Instead of applying a standard convolution, a depthwise convolution is used to process each channel separately, which helps reduce computational complexity. Another activation function is then applied. Finally, a second pointwise convolution projects the expanded channels back to match the input dimensions of the block.

Note that no activation function is added after the second pointwise layer, which, according to Sandler et al. [11], helps maintain representational power. Since we did not add a nonlinear activation after the second pointwise layer, our block follows a linear bottleneck. Notice that our block structure follows the same structure as the inverted residual block [11], see Fig. 3(c), in terms of increasing the channels and maintaining the representational power.

C. CHANNEL ATTENTION

Squeeze-and-Excitation (SE) [27] is the most popular channel attention that is used by many studies [22], [46]. As shown in Fig. 5, the SE block consists of two modules: the squeeze and the excitation. The squeeze module uses a global average pooling (GAP) to perform feature aggregation, and the excitation module contains two fully connected layers (FC). In addition, SE uses channel-wise multiplication to weight the channels of the feature maps.

According to SENet [27], the objective of the SE block is to provide a single parameter per channel to learn the importance of different channels. Hence, we incorporate the SE block after the depthwise layer to weight the expanding channels and extract the most important channels.

D. WIDEHRNET MODEL

This section provides a detailed explanation of the proposed model. We adopted the LiteHRNet model for its balance between accuracy and computational cost. However, as shown in Table. 1, the depthwise convolution layer in the CCW block of the LiteHRNet model has a greater accuracy impact than other layers. Therefore, we removed all layers except the depthwise layer in the CCW block. Then, we added a pointwise convolution before the depthwise convolution to expand the channels. Our expanding method follows the inverted residual block in terms of expanding and projecting the number of channels in a linear bottleneck block.

According to studies that build wide networks [9], [10], [11], [17], [48], increasing the channels led to redundant features. Hence, we applied channel expansion to part of the input channels, leaving the other part unchanged. To simplify, our proposed block starts by dividing the input channels into two groups. One half serves as an identity, while the other serves as input for the inverted residual block. The number of input channels for the second group is expanded by a factor of e. After expanding the channels using a pointwise layer, a deep convolution is used, which filters over a large number of channels at a low cost. Then, another pointwise convolution is utilized to restore the original number of channels. In addition, We followed a linear bottleneck, as in the inverted residual block, to preserve information from the loss. Finally, the channels of the identity group and the output of the inverted residual block are concatenated and shuffled.

To distinguish the important features among the massive number of channels, we added channel attention, specifically the SE block. This channel attention extracts the important feature by weighting all the expanding channels. The structure of our model is illustrated in Table. 2 and Fig. 4. The following section will provide a detailed discussion of the results obtained by different proposed blocks, i.e., the expanding block with and without channel attention.

TABLE 2. WideHRNet structure. The proposed model has the same depth as LiteHRNet_18. Hence, the depth of WideHRNet is set to 18.

Stage	Operator	#Output Channels	#Block	#Modules
Steam	shuffle block	32	1	1
$Stage_2$	proposed block fusion block	(40, 80) (40, 80)	2 1	2
$Stage_3$	proposed block fusion block	(40, 80,160) (40, 80,160)	2 1	4
$Stage_4$	proposed block fusion block	(40, 80,160, 320) (40, 80,160, 320)	2 1	2

IV. EXPERIMENTS

This section describes the experimental setup and its results. First, we will explain the dataset and the configuration values of the training and testing phases. Next, we will discuss the performance of the WideHRNet model. Finally, we show the results of tuning the expansion and reduction hyperparameters, as well as the effect of other lighter operations, including channel splitting, channel shuffling, and shortcut connection, on the accuracy of the model. The code of all experiments conducted in this paper is available to access[1]

[1] https://github.com/IsraaSamkari/WideHRNet

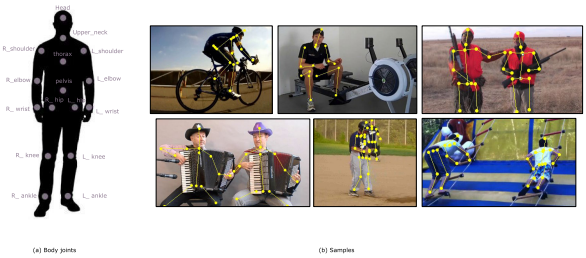


FIGURE 6. The MPII dataset. It provides (a) annotated 16 keypoints and (b) images with different human poses, including simple and complex poses.

A. IMPLEMENTATION DETAILS

In this section, we provide more detailed information about the dataset used to evaluate the WideHRNet model, as well as the configuration values for the training and validation phases.

1) DATASET AND EVALUATION METRIC

The MPII Human Pose dataset [49] contains about 25K images and 40K person pose instances, each with a label of 16 keypoints. We train our network on the MPII training set that contains 22K images and evaluate them on the MPII validation set that includes 3K images. The MPII images are from YouTube videos, and the pose in these images represents one of the human daily activities such as sitting, walking, running, dancing, bicycling, etc. Fig. 6 shows examples of the MPII images. The head-normalized Probability of the Correct Keypoint (PCKh) metric is used to evaluate the proposed model.

2) TRAINING

The WideHRNet model was trained on a single GeForce RTX 3060 GPU. For a fair comparison, we used LiteHRNet's default training setting. Hence, the optimizer is set to Adam with a base learning rate of 2e−3, batch size set to 32, epoch set to 210, and the image is resized to 256 × 256. The same goes for setting values of detection boxes. The human detection box aspect ratio is set to 4:3. In addition, several data augmentations are used, including random flipping, random scale, and random rotation.

3) TESTING

We follow the same testing setting as LiteHRNet. The LiteHRNet adopted a top-down paradigm in the test phase. This paradigm has two stages. The first stage is pose detection using the bounding box, which, for the MPII dataset, we used the bound boxes present in the annotation file. The second stage is to predict the position of joints by estimating heatmaps, each of which shows the probability of a particular joint being in a particular position in the image.

It is worth noting that we found that the MPII bounding boxes affect the final result because their boundaries include several other poses, not just the target pose. However, fixing the bounding box of person poses is beyond the scope of our research. Yet, we used an object detection model to demonstrate that the quality of the boundaries affects the accuracy of the pose estimation model. We have discussed this point in detail in Appendix.

TABLE 3. Evaluation results between WideHRNet model and the other models. These models are evaluated on the MPII val set, where the input size is 256 × 256. #Params and FLOPs indicate the model size and complexity, respectively. Pretrain: pretrain the backbone on ImageNet. Bold means the best result.

Model	Backbone	Pretrain	#Params (M)	FLOPs (G)	PCKh
Large Network					
[1]Stacked Hourglass [16]	Hourglass	N	26.0	55	90.9
HRNet_W32 [12]	HRNet_W32	Y	28.5	7.10	90.3
DARK [13]	HRNet_W32	Y	28.5	7.10	90.6
UPD [14]	HRNet_W32	Y	28.5	7.10	90.4
[1]ViTPose-H [33]	ViTPose-H	Y	637.2	122.8	94.1
Small Network					
MobileNetV2 [11]	MobileNetV2	N	9.6	1.97	85.4
MobileNetV3 [24]	MobileNetV3	N	8.7	1.8	84.3
ShuffleNetV2 [26]	ShuffleNetV2	N	7.6	1.70	82.8
Small HRNet [23]	HRNet-W18	N	1.3	0.7	80.2
LiteHRNet_18 [22]	LiteHRNet_18	N	**1.1**	**0.27**	86.1
LiteHRNet_30 [22]	LiteHRNet_30	N	1.8	0.42	87.0
RTMPose-m [44]	CSPNeXt-m	N	-	2.57	88.9
WideHRNet_18	WideHRNet_18	N	2.7	0.96	87.7
WideHRNet_18+SE	WideHRNet_18	N	4.4	0.97	**88.47**

TABLE 4. A comparison of how accurately each joint is estimated in the WideHRNet and LiteHRNet. The MPII val set is used for evaluation.

Model	Head	Should.	Elb.	Wrist	Hip	Knee	Ank.
LiteHRNet_18	95.8	93.8	85.2	79.1	86.7	80.2	75.2
LiteHRNet_30	96.5	94.7	87.1	81.3	87.2	81.6	77.5
WideHRNet_18	96.2	94.6	87.9	82.3	87.2	82.5	78.8
WideHRNet_18+SE	**96.6**	**95.3**	**88.6**	**83.6**	**88.1**	**83.6**	**79.7**

B. RESULTS

LiteHRNet provides a network that optimizes accuracy and efficiency by utilizing a CCW block. Nevertheless, as shown in Table. 1, analysis indicates that some operations within this block have low impact, except the depthwise layer. Consequently, we removed all operations within the CCW block and kept only the depthwise layer. Then, two pointwise layers were added before and after the depthwise layer to expand and project the number of channels, respectively, so the input and output can be added. We utilized this new block as a fundamental component in a high-resolution network, leading to a model with wider channels, which we have named WideHRNet.

The depth of WideHRNet is set to 18, which is the same as LiteHRNet_18. Then, we evaluated it on the MPII validation set. Table. 3 shows that the proposed model outperforms LiteHRNet_18 without adding significant computational complexity, as its value of GFLOPs is lower than 1. Compared to the CCW block, the proposed block significantly boosts accuracy to 87.7%. In addition, as shown in Table. 4, WideHRNet was able to predict the position of small joints, specifically the wrist and ankle, better than LiteHRNet. Fig. 7 shows the quality results between the LiteHRNet_18 and WideHRNet_18.

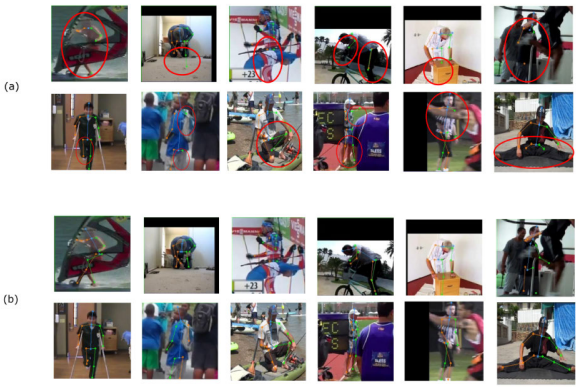
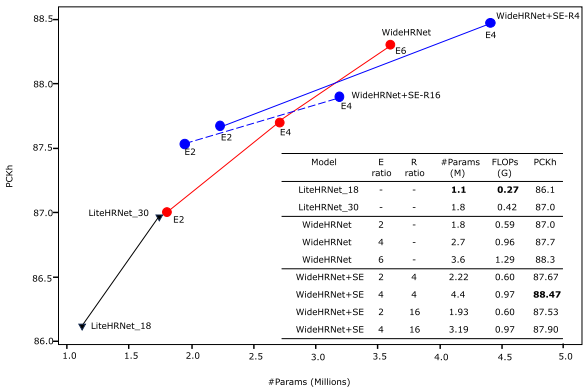


FIGURE 7. Evaluation quality results between (a) LiteHRNet_18 model and (b) the WideHRNet_18 model.



Model	E ratio	R ratio	#Params (M)	FLOPs (G)	PCKh
LiteHRNet_18	-	-	**1.1**	**0.27**	86.1
LiteHRNet_30	-	-	1.8	0.42	87.0
WideHRNet	2	-	1.8	0.59	87.0
WideHRNet	4	-	2.7	0.96	87.7
WideHRNet	6	-	3.6	1.29	88.3
WideHRNet+SE	2	4	2.22	0.60	87.67
WideHRNet+SE	4	4	4.4	0.97	**88.47**
WideHRNet+SE	2	16	1.93	0.60	87.53
WideHRNet+SE	4	16	3.19	0.97	87.90

FIGURE 8. Ablation study on the expansion and reduction ratios of the proposed block. E2, E4, and E6 mean expanding channels with different values (2, 4, and 6). R4 and R16 mean reduction ratios with different values (4 and 16).

It is important to note that the expanded channels in the proposed block have the same weights. To address this, the attention mechanism was added after the depthwise layer to reweight all the expanded channels. The attention mechanism used in this study is the SE block. As shown in Table.3 and Table.4, integrating the SE block into the WideHRNet significantly increased accuracy, reaching 88.47%, without adding complexity. However, this modification also led to doubling the number of parameters, which can be reduced by adjusting the reduction hyperparameter. Tuning hyperparameters will be discussed in Section IV-C.

Table. 3 presents a performance comparison between WideHRNet and other state-of-the-art models, including large and small networks. Notice that the WideHRNet with SE achieves good accuracy with few parameters and low FLOPs compared to the big networks. However, compared to small networks, like small HRNet and LiteHRNet_18, WideHRNet needs more parameters and FLOPs to get high accuracy. Nevertheless, the WideHRNet achieves the desired performance with fewer layers compared to lightweight networks that achieve high accuracy by making their networks deeper, e.g., LiteHRNet_30. Overall, the performance of the proposed model outperforms the existing efficient models such as MobileNetV3, ShuffleNetV2, and RTMPose.

C. ABLATION STUDY

In this section, we perform an ablation study on the proposed block to see how each component affects the performance. Hence, we studied the effect of increasing and decreasing the number of channels. In addition, we study the impact of channel split, channel shuffle, and shortcut connection on WideHRNet.

1) WIDE CHANNELS

The proposed block achieves high accuracy due to the increased number of channels. This increase is controlled by the e hyperparameter, which is set to 4 in Table. 3. In this section, the hyperparameter e is set to different values to measure its effect on the model accuracy. The expansion ratio was initially set to 2 and gradually doubled until we reached our hardware resource limit. As illustrated in Fig. 8, the higher the expansion ratio values, the more accurate the WideHRNet model is. However, this also increases the model's size and computational requirements. Yet, the most favorable result occurs when the expansion ratio is set to 6, resulting in a model accuracy of 88.3%. This result is close to WideHRNet with SE but with lower parameters and higher complexity.

Regarding the proposed block that includes SE, SE introduces more parameters, which can be controlled by changing the hyperparameter r (reduction ratio). Thus, we also tuned this hyperparameter to decrease the model size. Unfortunately, changing the hyperparameter r to different values reduced the model's accuracy, as shown in Fig. 8. In general, the adjustments in the expansion and reduction provide a balance between accuracy and efficiency. All these tuning results are shown in Fig. 8.

2) CHANNEL SPLIT

Increasing channels may lead to redundancy in filters [48] and thus increase unnecessary parameters in the model. To emphasize this point, we studied the effectiveness of dividing the input channels into groups. Previous experiments in Section IV-C1 have shown that dividing the input channels into two groups reduces the number of parameters while maintaining good accuracy (see Table. 3 and Fig. 8). However, it is unknown whether expanding the channels without splitting the input channels will increase the model accuracy. Therefore, in this section, the split channels were removed, and the inverted residual block was directly used. As a result, all input channels are expanded by the e ratio. The results are shown in Table. 5 indicate that model accuracy increased slightly, but this improvement came at the cost of model size and complexity. The number of parameters has increased significantly, and the complexity has tripled. This experiment demonstrated that applying the expanded layer without dividing the input channels introduces redundant

TABLE 5. Ablation study on the proposed block with and without using the channel split. The MPII val set is used for evaluation. SE: squeeze-excitation, E: expanding ratio of expansion layer, and R: reduction ratio of squeeze-excitation block.

Model	E ratio	R ratio	#Params (M)	FLOPs (G)	PCKh
Without Channel Split					
WideHRNet	4	-	7.89	2.83	88.50
WideHRNet+SE	4	4	14.68	2.85	88.90
With Channel Split					
WideHRNet	4	-	2.7	0.96	87.7
WideHRNet+SE	4	4	4.4	0.97	88.47

TABLE 6. Ablation study on the proposed block with and without using the channel shuffle. The MPII val set is used for evaluation. E: expanding ratio of expansion layer, and R: reduction ratio of squeeze-excitation (SE).

Model	E ratio	R ratio	#Params (M)	FLOPs (G)	PCKh
Without Channel Shuffle					
WideHRNet	4	-	2.7	0.96	87.7
WideHRNet	6	-	3.6	1.29	88.12
WideHRNet+SE	4	4	4.4	0.97	88.19
WideHRNet+SE	4	16	3.19	0.97	87.90
With Channel Shuffle					
WideHRNet	4	-	2.7	0.96	87.7
WideHRNet	6	-	3.6	1.29	88.3
WideHRNet+SE	4	4	4.4	0.97	88.47
WideHRNet+SE	4	16	3.19	0.97	87.90

TABLE 7. Ablation study on the proposed block with and without using the shortcut connection. The MPII val set is used for evaluation. E: expanding ratio and R: reduction ratio.

Model	E ratio	R ratio	#Params (M)	FLOPs (G)	PCKh
Without Shortcut Connection					
WideHRNet	4	-	2.7	0.96	87.65
WideHRNet	6	-	3.6	1.29	88.1
WideHRNet+SE	4	4	4.4	0.97	88.39
WideHRNet+SE	4	16	3.19	0.97	87.83
With Shortcut Connection					
WideHRNet	4	-	2.7	0.96	87.7
WideHRNet	6	-	3.6	1.29	88.3
WideHRNet+SE	4	4	4.4	0.97	88.47
WideHRNet+SE	4	16	3.19	0.97	87.90

TABLE 8. The ablation study of linear and nonlinear bottlenecks. A linear bottleneck means removing the activation function after the second pointwise layer, while a nonlinear bottleneck means adding the activation function.

Model	E ratio	R ratio	#Params (M)	FLOPs (G)	PCKh
Linear bottleneck					
WideHRNet	4	-	2.7	0.96	87.65
WideHRNet+SE	4	4	4.4	0.97	88.39
Nonlinear bottleneck					
WideHRNet	4	-	2.7	0.96	87.6
WideHRNet+SE	4	4	4.4	0.97	88.06

parameters to the model. In contrast, dividing the input channels before applying the expanded layer preserves the model's accuracy and efficiency.

3) CHANNEL SHUFFLE

As discussed in the previous section, splitting the input channels into two branches helps reduce model complexity while maintaining accuracy. However, the effect of using the channel shuffle operation after joining the channels from different branches is still unknown. Therefore, this section will discuss the impact of channel shuffle in the proposed block. ShuffleNetV2 [26] suggests using a channel shuffle to exchange information between branches. Our proposed block has two branches: one remains unchanged (identity), while the other performs convolutions. The output channels from these branches are then concating. Applying the channel shuffle after the branch concatenation should facilitate information exchange. However, as shown in Table. 6, the results indicate no big difference in accuracy when the channel shuffle is added or removed. Furthermore, removing the channel shuffle did not affect the model size or complexity. The conclusion from this experiment is that the channel shuffle operation can only improve performance by around 0.1 to 0.3 points.

4) SHORTCUT CONNECTION

Deep networks may face the vanishing gradient problem [11], where the early layers of the network stop learning due to the small values that come from backpropagation. The depth of the WideHRNet is 18, and to avoid the vanishing gradient problem in the WideHRNet, the proposed block uses a shortcut connection between thin layers. In this section, we conducted an ablation study on the shortcut connection and compared the effect of the proposed block with and without a shortcut connection on the model accuracy. As demonstrated in Table. 7, removing the shortcut connection from the proposed block did not lead to a vanishing gradient problem. However, using the shortcut connection in the proposed block is useful, as it slightly improves accuracy and does not add complexity to the model.

5) LINEAR AND NONLINEAR BOTTLENECK

In deep learning, the concept of a linear bottleneck refers to a design choice in neural network architecture where the dimensionality of data is reduced through a linear transformation without the use of activation functions. This technique is often employed to retain essential features of the input data [11]. In this section, we study the effect of adding and removing the ReLU activation function after the second pointwise layer in the proposed block. As shown in the Table. 8, the performance of the proposed block that follows a linear bottleneck is better than a nonlinear bottleneck, where the accuracy drops slightly when using the nonlinear bottleneck.

V. CONCLUSION

We proposed WideHRNet, a new efficient model for estimating human pose from the images. WideHRNet is built on the LiteHRNet_18 model, taking advantage of its high-resolution representation and parallel performance. The main contribution of WideHRNet is its basic block,

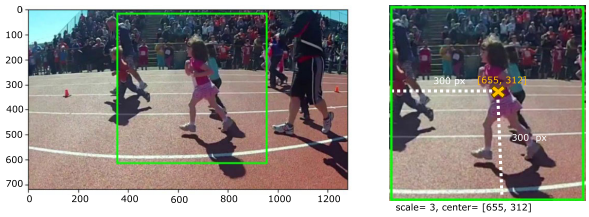


FIGURE 9. MPII Bounding Box (BBox): The scale and center indicate the values of the BBox length and center, respectively. To obtain the BBox length, the scale value must be multiplied by 200.

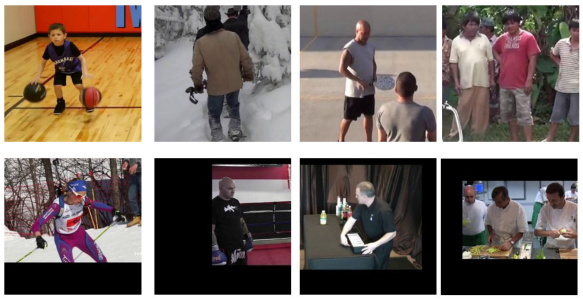


FIGURE 10. Samples from the MPII val set after cropping.

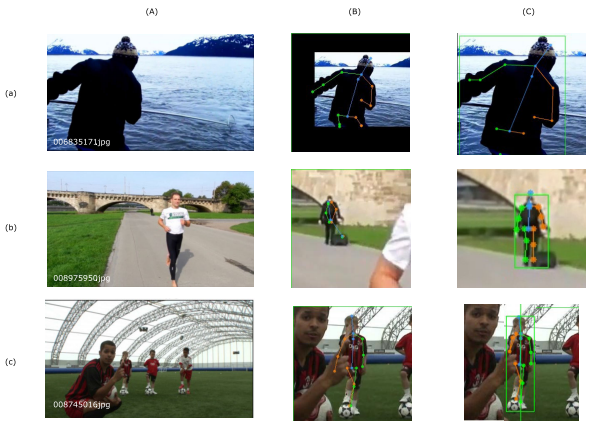


FIGURE 11. The evaluation of WideHRNet model using (A) MPII val set with (B) the ground truth MPII bounding box and (C) object detection model. The key scenarios include: (a) pose surrounded by black pixels, (b) small pose size, and (c) overlapping keypoints.

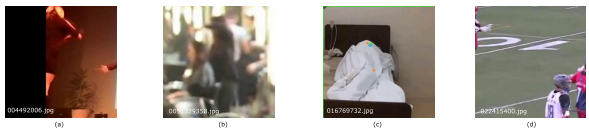


FIGURE 12. Examples of poor-quality images in the MPII dataset include some with (a) unclear poses, (b) blurring effects, (c) hidden poses, and (d) no pose centered in the image.

which increases the number of channels to capture various joint sizes. Besides expanding the channels, WideHRNet uses methods like channel splitting, shuffling, and attention techniques to boost efficiency. Additionally, the structure of the basic block follows the linear bottleneck to maintain the representational power. We show how each of these methods impacts the model's performance. By making the network wider and using the attention mechanism, the results on the MPII show that the WideHRNet outperforms the existing efficient models with an accuracy of 88.47% and a computational complexity of less than 1 GFLOP.

We proved that increasing the channels in each layer allows the model to estimate the small joint more accurately. However, we could not increase the number of channels beyond a certain limit due to hardware limitations. Therefore, many experiments are needed to find the appropriate expanding ratio. In addition, WideHRNet suffers from high parameters due to the use of the squeeze-excitation (attention block). Therefore, in future work, this block needs to be replaced with a free parameter block. Furthermore, regarding the MPII dataset, we noticed its bounding box is inaccurate, with each box including more than one person. We proved that fixing the bounding box of MPII will enhance the performance of the pose estimation model. However, fixing all the bounding box values is out of our scope. Hence, we left it as a future work.

APPENDIX
BOUNDING BOX OF MPII DATASET

This section will discuss the impact of the bounding box (BBox) on the results of the WideHRNet model. It is known that there are two approaches for estimating multipose [1], [2]: bottom-up and top-down. This paper follows the top-down approach that first identifies the person's location through a detection model and then predicts the body joints. However, the studies [12], [13], [14], [16], [22], [33], [41] that do the evaluation on MPII dataset, they use the BBox ground truth that provided by the MPII dataset. In our experiment, we did the same. Yet, we found one problem with this BBox.

The MPII dataset stores BBox values as BBox length and BBox center (see Fig. 9). However, we observed that the BBox often has a large boundary, introducing unnecessary information such as noisy backgrounds or including multipose. Additionally, if the person's pose is at the edge of the image, the remaining pixels inside the BBox are filled with black. Fig. 10 presents some samples from the validation set of the MPII dataset after extracting the BBox.

Despite the above problems, there are some samples in the MPII validation set that our model fails to evaluate accurately. This is due to the use of BBox present in the MPII dataset. To verify if the problem is from the MPII BBox, we used a Faster R-CNN [50] detection model. As shown in Fig. 11, with a detection model, our network is able to estimate the pose keypoints accurately. However, even by using a detector, we found some samples in the validation set hard to recognize, even with the human eye (see Fig. 12). However, we argue that if any model is able to estimate such samples accurately, it indicates that the model has low generalization.

Overall, in this section, we evaluated the quality of the MPII validation set and showed how the BBox accuracy affected the pose estimation result. However, fixing the values of the MPII BBox is beyond our scope. Therefore, we left it as an open area that needs to be solved.

REFERENCES

[1] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *ACM Comput. Surv.*, vol. 56, no. 1, pp. 1–37, 2023.

[2] E. Samkari, M. Arif, M. Alghamdi, and M. A. Al Ghamdi, "Human pose estimation using deep learning: A systematic literature review," *Mach. Learn. Knowl. Extraction*, vol. 5, no. 4, pp. 1612–1659, Nov. 2023.

[3] Y. Miao, J. Yang, B. Alzahrani, G. Lv, T. Alafif, A. Barnawi, and M. Chen, "Abnormal behavior learning based on edge computing toward a crowd monitoring system," *IEEE Netw.*, vol. 36, no. 3, pp. 90–96, May 2022.

[4] X. Chen, S. Kan, F. Zhang, Y. Cen, L. Zhang, and D. Zhang, "Multiscale spatial temporal attention graph convolution network for skeleton-based anomaly behavior detection," *J. Vis. Commun. Image Represent.*, vol. 90, Feb. 2023, Art. no. 103707.

[5] L. Kumarapu and P. Mukherjee, "AnimePose: Multi-person 3D pose estimation and animation," *Pattern Recognit. Lett.*, vol. 147, pp. 16–24, Jul. 2021.

[6] L. Lonini, Y. Moon, K. Embry, R. J. Cotton, K. McKenzie, S. Jenz, and A. Jayaraman, "Video-based pose estimation for gait analysis in stroke survivors during clinical assessments: A proof-of-concept study," *Digit. Biomarkers*, vol. 6, no. 1, pp. 9–18, Jan. 2022.

[7] B. Saunders, N. Cihan Camgoz, and R. Bowden, "Everybody sign now: Translating spoken language to photo realistic sign language video," 2020, *arXiv:2011.09846*.

[8] Y. Tian, H. Zhang, Y. Liu, and L. Wang, "Recovering 3D human mesh from monocular images: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15406–15425, Dec. 2023.

[9] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*.

[10] J. Yu, Y. Fan, J. Yang, N. Xu, Z. Wang, X. Wang, and T. Huang, "Wide activation for efficient and accurate image super-resolution," 2018, *arXiv:1808.08718*.

[11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[12] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5693–5703.

[13] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7093–7102.

[14] J. Huang, Z. Zhu, F. Guo, and G. Huang, "The devil is in the details: Delving into unbiased data processing for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5700–5709.

[15] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "HRFormer: High-resolution vision transformer for dense predict," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 7281–7293.

[16] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 483–499.

[17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[18] D. Qin, C. Leichner, M. Delakis, M. Fornoni, S. Luo, F. Yang, W. Wang, C. Banbury, C. Ye, B. Akin, V. Aggarwal, T. Zhu, D. Moro, and A. Howard, "MobileNetV4—Universal models for the mobile ecosystem," 2024, *arXiv:2404.10518*.

[19] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, and G. Anderson, "Wide & deep learning for recommender systems," in *Proc. 1st Workshop Deep Learn. Recommender Syst.*, 2016, pp. 7–10.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[21] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[22] C. Yu, B. Xiao, C. Gao, L. Yuan, L. Zhang, N. Sang, and J. Wang, "Lite-HRNet: A lightweight high-resolution network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10440–10450.

[23] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[24] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

[25] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[26] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 116–131.

[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[28] M. A. Khan, "Multiresolution coding of motion capture data for real-time multimedia applications," *Multimedia Tools Appl.*, vol. 76, no. 15, pp. 16683–16698, Aug. 2017.

[29] T. Alsubait, T. Sindi, and H. Alhakami, "Classification of the human protein atlas single cell using deep learning," *Appl. Sci.*, vol. 12, no. 22, p. 11587, Nov. 2022.

[30] M. K. Shambour and A. Gutub, "Progress of IoT research technologies and applications serving Hajj and umrah," *Arabian J. Sci. Eng.*, vol. 47, no. 2, pp. 1253–1273, Feb. 2022.

[31] E. A. Khan and M. K. Y. Shambour, "An analytical study of mobile applications for Hajj and umrah services," *Appl. Comput. Informat.*, vol. 14, no. 1, pp. 37–47, Jan. 2018.

[32] T. Alafif, A. Hadi, M. Allahyani, B. Alzahrani, A. Alhothali, R. Alotaibi, and A. Barnawi, "Hybrid classifiers for spatio-temporal abnormal behavior detection, tracking, and recognition in massive Hajj crowds," *Electronics*, vol. 12, no. 5, p. 1165, Feb. 2023.

[33] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose: Simple vision transformer baselines for human pose estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 38571–38584.

[34] G. Hua, L. Li, and S. Liu, "Multipath affinage stacked—Hourglass networks for human pose estimation," *Frontiers Comput. Sci.*, vol. 14, pp. 1–12, Jan. 2020.

[35] T. Xu and W. Takano, "Graph stacked hourglass networks for 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16105–16114.

[36] Y. Li, R. Liu, X. Wang, and R. Wang, "Human pose estimation based on lightweight basicblock," *Mach. Vis. Appl.*, vol. 34, no. 1, p. 3, Jan. 2023.

[37] R. Li, A. Yan, S. Yang, D. He, X. Zeng, and H. Liu, "Human pose estimation based on efficient and lightweight high-resolution network (EL-HRNet)," *Sensors*, vol. 24, no. 2, p. 396, Jan. 2024.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[40] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose++: Vision transformer for generic body pose estimation," 2022, *arXiv:2212.04246*.

[41] Q. Li, Z. Zhang, F. Xiao, F. Zhang, and B. Bhanu, "Dite-HRNet: Dynamic lightweight high-resolution network for human pose estimation," 2022, *arXiv:2204.10762*.

[42] L. Rui, Y. Gao, and H. Ren, "EDite-HRNet: Enhanced dynamic lightweight high-resolution network for human pose estimation," *IEEE Access*, vol. 11, pp. 95948–95957, 2023.

[43] Z. Li, M. Xue, Y. Cui, B. Liu, R. Fu, H. Chen, and F. Ju, "Lightweight 2D human pose estimation based on joint channel coordinate attention mechanism," *Electronics*, vol. 13, no. 1, p. 143, Dec. 2023.

[44] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, and K. Chen, "RTMPose: Real-time multi-person pose estimation based on MMPose," 2023, *arXiv:2303.07399*.

[45] R. Li, H. Huang, and Y. Zheng, "Human pose estimation based on lite HRNet with coordinate attention," in *Proc. 7th Int. Conf. Intell. Comput. Signal Process. (ICSP)*, Apr. 2022, pp. 1166–1170.

[46] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11534–11542.
[47] M. H. Guo, T. X. Xu, J. J. Liu, Z. N. Liu, P. T. Jiang, T. J. Mu, S. H. Zhang, R. R. Martin, M. M. Cheng, and S. M. Hu, "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, 2022.
[48] D. Zhou, M. Ye, C. Chen, T. Meng, M. Tan, X. Song, Q. Le, Q. Liu, and D. Schuurmans, "Go wide, then narrow: Efficient training of deep thin networks," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 11546–11555.
[49] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.
[50] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

ESRAA SAMKARI received the bachelor's degree (Hons.) in computer science from Umm Al-Qura University, Makkah, Saudi Arabia, in 2019, where she is currently pursuing the master's degree. Her research interests include artificial intelligence and computer vision.

MUHAMMAD ARIF received the Bachelor of Engineering degree from Ned University, Karachi, Pakistan, in 1990, the M.S. degree from Quid-e-Azam University, Islamabad, Pakistan, in 1993, and the Ph.D. degree in system information science from Tohoku University, Japan, in 1999. He completed two years of a Postdoctoral Fellowship in intelligent systems with Tohoku University, in 2003. Currently, he is a Professor with the Department of Computer Science and Artificial Intelligence, College of Computing, Umm Al-Qura University, Saudi Arabia. He has published more than 150 papers in various journals and conference proceedings. His research interests include artificial intelligence, pattern recognition, evolutionary computing, biometrics, and biomedical signal and image processing.

MANAL ALGHAMDI received the Ph.D. degree in computer vision from The University of Sheffield, U.K., in 2015. Her study involved video representation and video similarity measurements. In 2019, she was a Visiting Scholar with the Department of Electrical and Computer Engineering, University of Miami, USA, where she developed an interest in deep learning and its application in healthcare applications. She is currently an Associate Professor with the Department of Computer Science and Artificial Intelligence, Umm Al-Qura University (UQU), Saudi Arabia. She focuses on developing and evaluating video and image processing techniques for various applications, including video representation similarity measurements and crowed analysis. Her research interests include machine learning, computer vision, and security.

MOHAMMED A. AL GHAMDI (Member, IEEE) received the bachelor's degree (Hons.) in computer science from King Abdul Aziz University, Jeddah, Saudi Arabia, in 2004, the master's degree (Hons.) in internet software systems from Birmingham University, Birmingham, U.K., in 2007, and the Ph.D. degree in computer science from the University of Warwick, U.K., in 2012. Since 2012, he has been with the Department of Computer Science, Umm Al-Qura University, Makkah, Saudi Arabia, as an Assistant Professor and an Associate Professor. He is currently a Full Professor of computer science and artificial intelligence and the Founder of the Scientific Chair of data and artificial intelligence with Umm Al-Qura University. He has authored more than 50 papers in international conferences and journals, such as IEEE Systems Journal, IEEE Transactions on Engineering Management, IEEE Access, IEEE International Conference on Scalable Computing and Communications, and International Conference on Cloud Computing and Services Science. His research interests include machine learning, data analysis, AI, cloud computing, and cybersecurity.