BK-POSE: A LIGHTWEIGHT MODEL FOR MULTI-PERSON POSE ESTIMATION IN THE WILD

Van Giang Nguyen[1,*], *Chan Hung Nguyen*[1], *Quang Dich Nguyen*[1], *Cong Dong Trinh*[1]

DOI: 10.56651/lqdtu.jst.v12.n1.655.ict

Abstract

Human pose estimation is an essential topic in computer vision research, which has numerous applications in various fields. In this article, we propose a lightweight deep learning architecture called BK-Pose for recognizing human keypoints from images and videos that can be executed on edge computing devices. The main contributions of our work are two folds: (1) A lightweight bottom-up deep learning model is introduced that can be deployed on edge devices and can achieve high frames-per-second (FPS) rates while maintaining acceptable accuracy in suitable environments. (2) The incorporation of a novel focal $\mathbb{L}2$ loss ($\mathbb{FL}$) technique allows for the effective balancing between "hard" and "easy" keypoints samples during training.

The performance of the BK-Pose model is evaluated within a classroom environment for capturing students' keypoints and demonstrate its efficacy. Our results show promise for further research in activity recognition using the proposed architecture.

Index terms

Human pose estimation, pose estimation, deep learning, convolutional neural network.

1. Introduction

Human Pose Estimation (HPE) is a significant area of research in computer vision, which involves the estimation of human body parts configuration using input data from sensors, such as images and videos [1]. HPE provides geometric and motion information on the human body, which finds applications in various fields, such as human-computer interaction, human activity recognition, augmented reality (AR), virtual reality (VR), violence detection, healthcare, and more. Over the last few years, deep learning solutions have emerged as a dominant approach, surpassing traditional computer vision methods in many tasks, such as image classification, semantic segmentation, object detection, and more.

[1] ICEA Institute of Control Engineering and Automation, HUST

[*] Corresponding author, email: giangnv.soict.hust@gmail.com

Deep learning algorithms have already made significant progress and achieved outstanding results in HPE workloads. However, issues including occlusion, inadequate training data, and depth ambiguity remain obstacles to overcome. 2D HPE from pictures and movies with 2D pose annotations is simple to do, and great performance has been achieved utilizing deep learning approaches for human pose estimate of a single person.

In recent years, there has been a growing interest among scholars to address the challenge of highly occluded multi-person HPE in complex environments. However, obtaining accurate 3D posture annotations in 3D HPE is more intricate than in 2D HPE. While motion capture devices can obtain 3D posture annotations in controlled laboratory settings, their use is limited in real-world scenarios. The primary challenge in 3D HPE from monocular RGB photos and videos is the presence of depth ambiguity. In addition, multi-view setups present a fundamental issue of viewpoint affiliation. To tackle these challenges, some studies have explored the use of sensors such as depth sensors, inertial measurement units (IMUs), and radio frequency devices. Nonetheless, these methods are often expensive and require specialized equipment.

Despite the ongoing research efforts in the field of HPE, the current models suffer from significant limitations. One of the primary issues is the high complexity of the models, which require substantial computing resources, such as Geforce RTX 2080, RTX 3060Ti, among others. The average HPE models contain a parameter range of 8M to 70M, which poses challenges in deploying them on a large scale, especially in natural environments. As a result, there is a need for marginal computing to reduce costs and ensure real-time processing of requests.

This research article proposes a novel architecture for HPE, named BK-Pose. The primary objective of BK-Pose is to generate high-resolution and high-quality heatmaps by incorporating a lightweight architecture, the BK-Pose block, in multiple layers of the network during the training process. The proposed approach also incorporates a "spatial attention mechanism" to supervise the network at smaller sizes and gather multi-scale hidden information for each keypoint and body component. This technique enhances the network's ability to refine the location of keypoints in future high-resolution layers using low-resolution heatmaps, resulting in the generation of heatmaps with increasing quality and resolution. The main contributions of this article can be summarized as follows:

- We introduce a lightweight bottom-up deep learning model that can be deployed on edge devices and can achieve high FPS rates while maintaining acceptable accuracy in suitable environments.
- We incorporate a novel focal $\mathbb{L}2$ loss technique to balance between "hard" and "easy" keypoints samples during training.

The rest of the article is organized as follows. Section 2 briefly reviews the previous works in applying deep learning to HPE problem. Section 3 describes the proposed method. The tested datasets and experimental configurations are provided in section 4. Section 5 presents experimental results and the analysis. The conclusions and future work are discussed in section 6.

2. Background and related work

Multi-person HPE is considerably more complex than for a single person since it involves determining the number of individuals present in the image, their locations, and how to organize the keypoints for each individual. To address these challenges, two main approaches have been developed for multi-person HPE, namely top-down and bottom-up methods. Top-down approaches typically employ off-the-shelf person detectors to extract a collection of bounding boxes (each corresponding to an individual) from the input images and then apply single-person pose estimators to each individual bounding box to produce the final multi-person poses.

In contrast to top-down approaches, bottom-up methods detect all body joints in an image and group them to individuals, while top-down approaches require individual identification for each person, making bottom-up methods faster.

2.1. Top-down pipeline

The top-down pipeline for multi-person HPE comprises two major components: a human body detector to generate bounding boxes for individuals and a single-person pose estimator to predict the positions of keypoints within these bounding boxes. Prior works such as [2]–[4] have focused on improving the modules in HPE networks. For instance, Xiao et al. [5] introduced deconvolutional layers to the ResNet backbone network to create a simple yet effective structure that generates high-resolution heatmaps, aiming to explore the fundamental question of how effective a basic approach can be for HPE. Similarly, Sun et al. [4] proposed the High-Resolution Net (HRNet) that integrates multi-resolution subnetworks and conducts multiple multi-scale fusions to produce accurate high-resolution representations.

Wang et al. [6] introduced a novel approach called GraphPCNN to enhance the accuracy of keypoint localization. This method is based on a two-stage graph-based and model-agnostic process that involves a localization subnet to obtain approximate keypoint locations and a graph pose refinement module to generate refined keypoint localization representations. On the other hand, Cai et al. [7] proposed a multi-stage network, which includes a Residual Steps Network (RSN) module for the efficient fusion of intra-level features to learn delicate local representations, and a Pose Refine Machine (PRM) module that aims to balance local and global representations in the features to attain more precise keypoint localization.

In multi-person scenarios, estimating postures under occlusion and truncation situations is a common challenge due to the overlapping nature of limbs. The initial phase of a top-down pipeline using human detectors may fail under such conditions, emphasizing the significance of occlusion or truncation resistance in multi-person HPE techniques. To address this issue, Iqbal and Gall [8] proposed a posture estimator based on a neural pose machine approach that estimates joint candidates with improved robustness under occlusion or truncation scenarios.

In the realm of HPE, occlusions and complex scenarios present formidable challenges for existing methodologies. To surmount these hurdles, recent works have introduced novel techniques for joint-to-person connection and multi-person pose estimation. Fang et al. [9] propose a regional multi-person pose estimation (RMPE) framework that partitions the pose estimation process into three key stages. First, a Symmetric Spatial Transformer Network is utilized to identify single person regions within imprecise bounding boxes. Second, the Parametric Pose Non-Maximum-Suppression algorithm addresses the problem of redundant detection. Finally, the Pose-Guided Proposals Generator augments training data to further enhance the performance of the model. Notably, the authors employ integer linear programming (ILP) to solve the joint-to-person connection problem, which remains a persistent challenge in HPE. The efficacy of the proposed RMPE framework is demonstrated through experiments on a diverse range of challenging datasets.

2.2. Bottom-up pipeline

In the context of HPE, the bottom-up pipeline is composed of two principal stages: body joint detection, which entails the extraction of local features and prediction of potential joint candidates, and joint candidate assembling, which involves grouping these candidates and constructing the final pose representation using part association strategies. This paradigm has been employed in various state-of-the-art methods, such as DeepCut [10], DeeperCut [11], and Realtime Multi-Person Pose Estimation [12]. These works have demonstrated the effectiveness of the bottom-up approach in accurately estimating the complex and articulated poses of multiple individuals in real-world scenarios.

Pishchulin et al. [10] proposed DeepCut, a two-step bottom-up approach for HPE, which relies on a Fast R-CNN based body part detector. In the first stage, DeepCut employs this detector to recognize all possible body components. The second stage utilizes an ILP framework to label each component and assemble them into a final posture representation. DeepCut was among the early bottom-up methods to achieve state-of-the-art performance in HPE. Its success demonstrates the viability of the two-step bottom-up approach for accurately estimating complex and articulated poses of multiple individuals in diverse scenarios.

The DeepCut method, while achieving high accuracy in multi-person HPE, is computationally expensive. To address this limitation, Insafutdinov et al. [11] proposed DeeperCut, which integrates a more powerful body part detector with an improved incremental optimization algorithm that leverages image-conditioned pairwise terms to group body parts. This results in enhanced performance and reduced computational complexity. Subsequently, Cao et al. [12] introduced the OpenPose framework, which employs heatmaps to predict the locations of keypoints and Part Affinity Fields (PAFs) to connect the keypoints to individual people. The use of PAFs facilitates efficient multi-person pose estimation, enabling OpenPose to achieve real-time performance. To further enhance the performance of OpenPose, Zhu et al. [13] incorporated redundant edges into the PAFs, thereby improving the connections between joints and yielding even higher accuracy than the baseline OpenPose framework. These advancements have significantly accelerated the pace of bottom-up multi-person HPE, making it a more practical and viable solution for diverse applications.

While OpenPose-based approaches have demonstrated excellent performance on high-resolution images, they face significant challenges in handling low-resolution images and scenes with occlusions. To address these limitations, Kreiss et al. [14] introduced PifPaf, a novel bottom-up approach that leverages a Part Intensity Field (PIF) to predict body part locations and a Part Association Field (PAF) to express joint associations. By incorporating these fields, PifPaf outperforms previous OpenPose-based techniques in low-resolution and occluded scenes. These advancements have the potential to significantly improve the accuracy and robustness of HPE in challenging real-world scenarios.

Bottom-up HPE techniques have shown promising results in recent years. Multi-task learning architectures have been increasingly used to integrate different components of the HPE pipeline. PersonLab, a multi-task model proposed by Papandreou et al. [15], combines posture estimation and person segmentation modules for keypoint detection and association. PersonLab employs short-range offsets to improve heatmaps, mid-range offsets to anticipate keypoints, and long-range offsets to group keypoints into instances. Another notable example of multi-task learning for HPE is MultiPoseNet, a model proposed by Kocabas et al. [16]. MultiPoseNet uses a pose residual network to perform simultaneous tasks of keypoint prediction, person identification, and semantic segmentation. These approaches show promising results and pave the way for future research in the field of HPE.

This article proposes a novel model for HPE that differs from previous approaches in several key aspects. Firstly, our model inherits the Blaze Block architecture [17] into the BK-Pose Block at multiple layers to achieve a balance between high and low-level features. To this end, skip connections are employed extensively across all levels of the network. Additionally, the Blaze Block architecture allows for the reduction of the network's parameter count, thereby enabling deployment to resource-constrained devices with faster computational capabilities. Secondly, the BK-Pose block is utilized to generate high-resolution and high-quality heatmaps, which provide multi-scale hidden information on each keypoint and body component by supervising at lesser sizes. The low-resolution heatmaps can facilitate location refinement in subsequent high-resolution layers, thus improving the quality and resolution of the generated heatmaps. Thirdly, we introduce a novel focal $\mathbb{L}2$ loss technique that balances between "hard" and "easy" keypoints samples. This technique aids in improving the model's overall performance in accurately estimating human poses.

3. The proposed method

3.1. Overall architecture

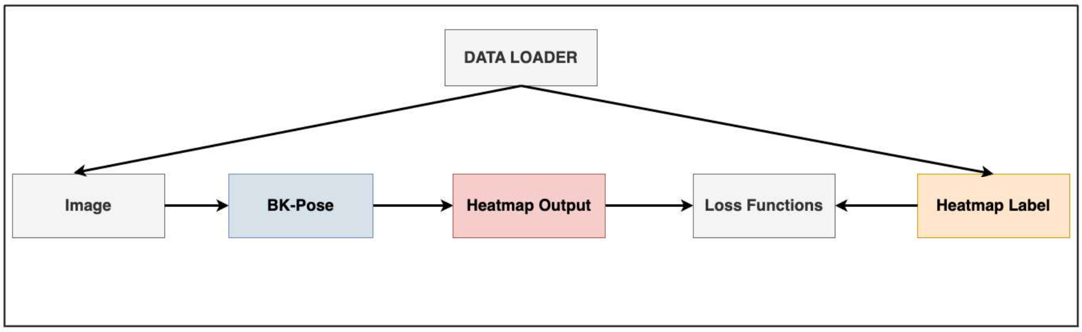

Fig. 1. BK-Pose Overall Architecture.

The article presents the system architecture, illustrated in Fig.1, which serves to delineate the training process. The training process incorporates a series of modules, as explicated below:

- The preprocessing data module: This module is responsible for the preparation of data for input into the model. Specifically, it performs data preprocessing operations to generate a tuple in the form (X, Y).
  - X is the preprocessed and normalized image.
  - Y is the label that includes the keypoint heatmap and the body part heatmap.
- BK-Pose: Operate on image data as input and generate a keypoint heatmap along with a heatmap corresponding to each body component.
- Loss function module: The loss function will control the error of the prediction result and the label heatmap.

3.2. Network structure

The BK-Pose architecture (Fig.2) has been constructed using an iterative encoder and decoder architecture inference to capture the diverse spatial extent of each keypoint in addition to their associations with other keypoints.

In this study, we adopt the Residual module [18] as the backbone, which is a widely-used architecture for feature extraction from images prior to inputting them into the network. The Residual backbone includes two Residual blocks that are simple yet effective, thereby serving as an ideal candidate for integration with the BK-Pose module.

The green hourglass depicts the design of stack BK-Pose modules. The down-sampling approach decreases the input feature map's spatial extent by half while increasing N in feature map channels C ($C = 64$ and $N = 32$). As the result, in the last stack of our BK-Pose module, the feature map has 192 channels.

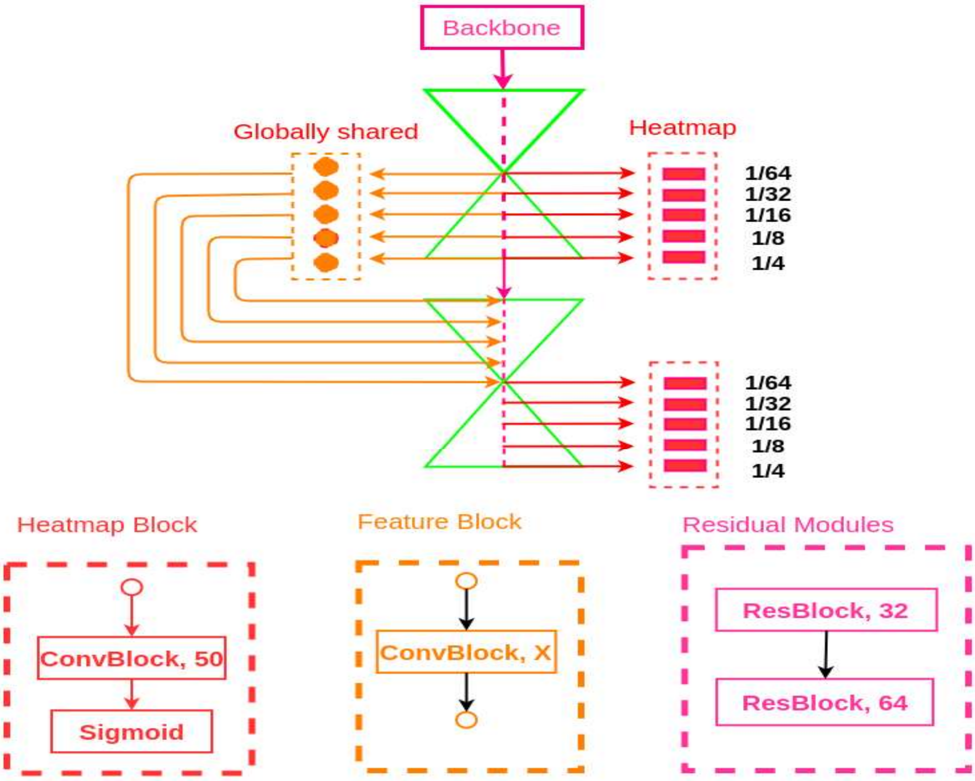

Fig. 2. Network structure: The pink dashed box refers to Backbone (Residual Modules). The feature maps with 5 *different scales those surrounded by the orange dashed box are extracted from each (stage) BK-Pose module (green hourglass). The output heatmaps are also generated by BK-Pose module.*

The present work leverages multi-scale supervision [19] to train the two stacked BK-Pose modules, generating heatmaps across 5 scales during training, with the addition of a "spatial attention mechanism". By supervising at lower sizes, the network is compelled to accumulate multi-scale hidden information for every keypoint and body component. The low-resolution heatmaps obtained assist in refining the location information in subsequent high-resolution layers, thus enhancing the quality and resolution of the heatmaps. Moreover, adaptive average pooling is employed to downsize the ground truth of keypoint and body part heatmaps to fractional sizes from the full-size base.

As the feature map at a given scale stores pose structural information, it exhibits strong self-correlation and is employed to train heatmaps at the same scale. This feature map not only stores pertinent features but also contributes to the 2^{nd} stack of BK-Pose.

3.3. BK-Pose block

The necessity to capture information at multiple scales motivates the design of the BK-Pose architecture. While local features such as limbs are critical for characterizing body posture, a comprehensive understanding of the entire body is necessary for HPE.

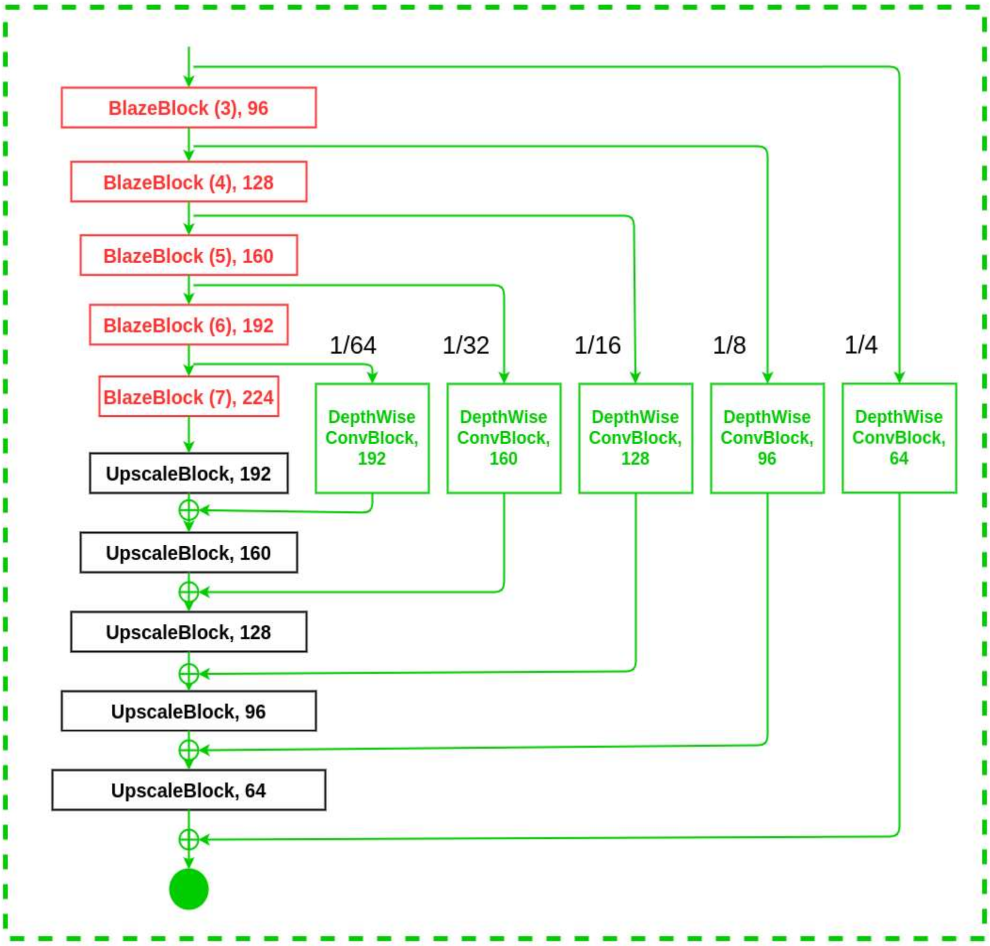

Fig. 3. The BK-Pose block with 5 scales downsizes image by half while increasing 32 *channels in down-sampling approach.*

To achieve this goal, the network must process information across different scales. One common approach is to use separate pipelines that analyze the image at different resolutions and subsequently aggregate features [20]. In contrast, we use multi-scale supervision to infer heatmaps at five different resolutions during training, which enables the model to capture structural information of each keypoint and body component across multiple scales.

The BK-Pose architecture is built upon Blaze Block [17], which downsamples the feature map to a low resolution before applying *DepthwiseConv*2*D* and *Convolution* 2*D* operations at the original pre-pooled resolution (Fig.3). After reaching the smallest resolution, the network uses upsampling and feature combining to decode the features. Specifically, we use the Tompson et al. [17] method to upsample the feature map to the closest neighbor of the lower resolution, and then we perform elementwise sum to integrate information from two nearby resolutions.

To further enhance the network's ability to capture global and local information, we stack multiple BK-Pose modules and use the output feature map of one module as the input to the next. By repeatedly conducting encoding and decoding inference, the model can reconsider the original predictions and features throughout the entire image. Crucially, intermediate heatmap predictions are used to guide the optimization process, which encourages the network to generate accurate predictions at every scale.

The technique of switching between scales is particularly important for structured tasks like HPE. Preserving the spatial position of features is critical for the final localization phase, as the precise location of a joint is essential for accurate pose estimation. By moving back and forth between different scales, the network can retain local information while examining and rethinking the overall coherence of the features, which is essential for avoiding contradictory evidence and anatomic impossibility.

3.4. Loss functions

The $\mathbb{L}2$ loss is frequently used to measure the distance between the predicted heatmaps and the target heatmaps, e.g. Cao et al. 2017 [19]; Ke et al. 2018 [21]. To deal with "hard" keypoints, the research Chen et al. 2018 [22] suggests $\mathbb{L}2$ loss which makes online "hard" keypoint extraction. We have suggested a novel loss, that is named *focal* $\mathbb{L}2$ loss, under the combined concept of keypoint and body part heatmaps, to address the two types of sample imbalance problems. The network will generate K keypoint heatmaps as well as B body connection part heatmaps. They will be created at 5 different scales at each stage of the stacking model.

In each generated heatmaps, the pixel value represent for the probability of which category the keypoint or body part belong to. Assume that, at stage n, the predicted score maps have a size of $w_i \times h_i$ are $S^n = (S_1^n, S_2^n, \cdots, S_{K+B}^n)$, i is the scale order, $n \in \{1, 2, \cdots, N\}$, where N is the number of stack BK-Pose. Assume that the ground truth heatmaps has the equal size of $S^* = (S_1^*, S_2^*, \cdots, S_{K+B}^*)$ and G is the Gaussian peak generation function. We have the ground truth score $S_j^*(p)$ at the pixel location $p(x, y) \in \mathbb{R}^{w_i \times h_i}$ in the $j - th$ heatmap, the formula of $S_j^*(p)$ is expressed as:

$$S_j^*(p) = \begin{cases} G(x, y|\mathbb{R}, \sigma_k, r_0), & 1 \leq j \leq K \\ G(x, y|\mathbb{R}, \sigma_b, d_0), & K < j \leq K + B \end{cases} \tag{1}$$

The expression of $Sd_j^n(p)$:

$$Sd_j^n(p) = \begin{cases} S_j^n(p) - \alpha, & S_j^*(p) > thre \\ 1 - S_j^n(p) - \beta, & otherwise \end{cases} \tag{2}$$

where σ_k and σ_b are the standard deviations of Gaussian peaks heatmap: keypoint and body part. The hyper-parameters r_0 and d_0 determine the boundaries of the ground truth Gaussian peaks. α, β are adjustment variables that are used to decrease the penalty of easy samples (both easy foreground and easy background pixels), allowing us to fully use the training data. The threshold for differentiating between foreground and background heatmap is $thre$.

The focal $\mathbb{L}2$ loss ($\mathbb{FL}$) at stage n, between the predicted heatmaps and target heatmaps of size $w_i \times h_i$ is calculated as follows:

$$\mathbb{FL}_i^n = \sum_{j=1}^{K+B} \sum_{p \in R^{w_i \times h_i}} [\eta \cdot I(j \leq K) + 1] \cdot W(p) \cdot |S_j^n(p) - S_j^*(p)| \cdot (1 - Sd_j^n(p))^\gamma \quad (3)$$

where W is the binary masking, $W(p) = 0$ since the annotation is hidden at the position p. We have the indicator function I, and the hyper-parameter η is used to equalize the keypoint heatmap loss with the body component heatmap loss. The term "scaling factor" is introduced $(1 - Sd_j^n(p))^\gamma$ presupposes two pieces of prior knowledge: The heatmap score of easy foregrounds are generally high (near to 1, e.g. 0.9); The heatmap score of easy backgrounds are generally low (typically less than 0.01 in practice). As a result, it has the potential to make simple samples less effective and contribute less during training, as focal loss suggests [23].

To make the balance in the gradient between the foreground and background value of pixels, we have created the parameter setting $\sigma_k = 9, \sigma_b = 7, thre = 0.01$ in this work, and we set $\eta = 2$ appropriately. Furthermore, we approximately set $\alpha = 0.1$ and $\beta = 0.02$. We suggest that they should be put close to 0 and $0 \leq \beta \leq \alpha$. More explanations of the main hyper-parameters are provided for a better understanding.

In our experiment creating heatmap, if we set the standard deviations of Gaussian peaks heatmap: keypoint and body part i.e., σ_k and σ_b, too small, the correct positioning information is retained, but the output heatmap at these peaks are low, giving in more false negatives. However, if these parameters are too large, the Gaussian peaks stretch out too much at inference time, the localization knowledge becomes fuzzy, decreasing localization accuracy (offset regression may help here). We have also used the $thre$ hyper-parameter and been being set at 0.01 to reduce the loss of numerous simple backdrop pixels.

The total loss of the stacked BK-Pose across 5 different scales can be written as:

$$L = \sum_{n=1}^{N} \sum_{i=1}^{5} \lambda_i^n \cdot \mathbb{FL}_i^n \quad (4)$$

in which $\lambda_1^n = 0.1, \lambda_2^n = 0.2, \lambda_3^n = 0.4, \lambda_4^n = 1.6$ and $\lambda_5^n = 6.4$ are used to balance between losses at 5 scales. In experiments, we find that by setting these values achieves the best performance on COCO dataset. These coefficients were chosen by experiments with 142 trials. Due to the scope of the article, this data is not presented here.

4. Experimental settings

This section presents the datasets used in the experiments and the parameter's setting for the tested models.

4.1. Datasets

In this article, we use the MS-COCO dataset and our keypoint dataset (BK-SAD) for training and evaluation.

Over than 200000 pictures as well as 250000 person instances were labeled using 17 keypoints in the COCO which contains train dataset, validation, and test dataset as described in table 1. Annotations on train and val (with over 150000 people and 1.7 million labeled keypoints) are publicly available. All the experiments in this project are trained only on the train set. Fig.4 presents some samples of the MS-COCO dataset.

Table 1. Number of image in each type of COCO dataset

Type of Dataset	Number of image
train2017	118287
val2017	5000
test2017	40670

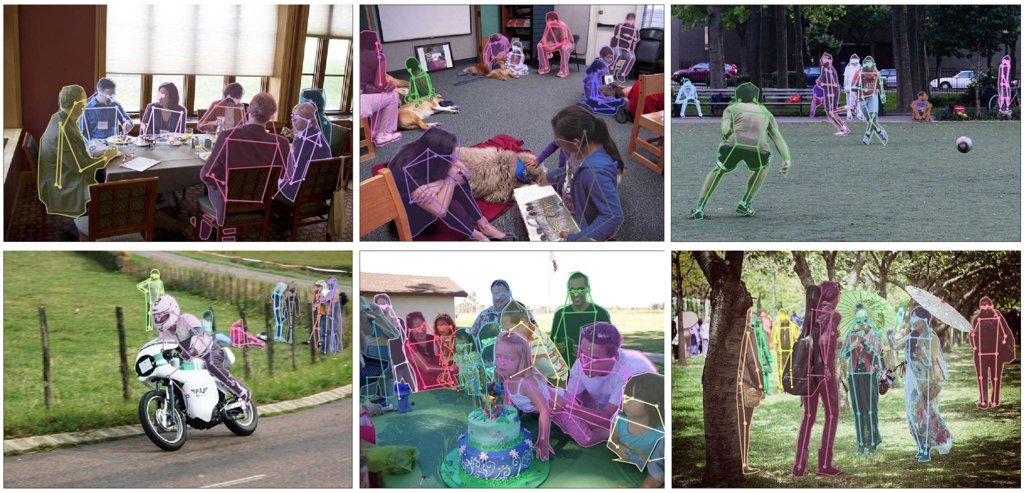

Fig. 4. COCO Dataset example. [1]

Differ from MS-COCO dataset, which covers all general activities, BK-SAD dataset focuses on classroom environment. It contains over 48,000 labeled samples of students raising their hands, over 34,000 labeled samples of students dozing off during class, and over 170,000 labeled samples of normal activities collected from several schools in Hanoi city as described in table 2. Fig.5 presents some samples of the BK-SAD dataset.

[1] https://cocodataset.org/#keypoints-2017

Table 2. Number of image in each activity type of BK-SAD dataset

Class Label	Train Images	Test Images	% of Total Dataset
Doze	24333	10429	13.62%
Hand-raising	39920	14537	19.00%
Normal	120358	51582	67.38%
Total	178611	76548	100%

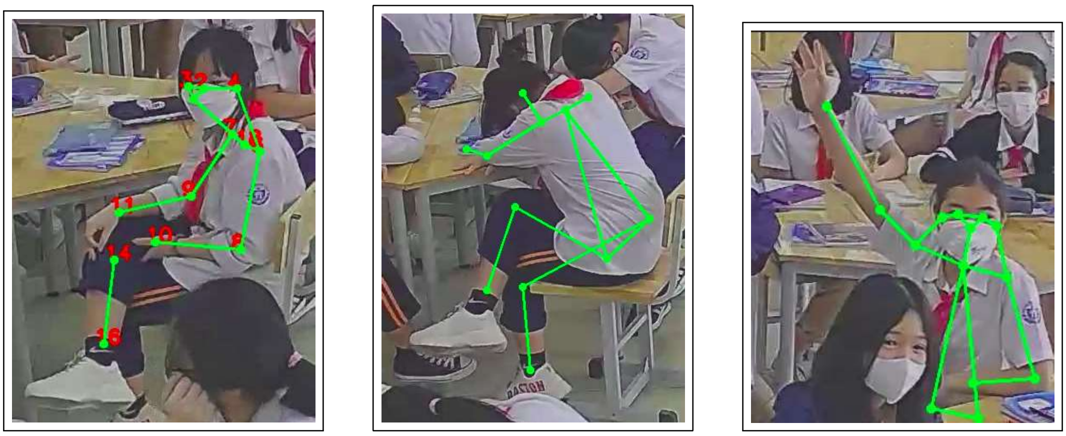

Fig. 5. BK-SAD Dataset example.

4.2. Data preparation

In this article, we perform the data preparation steps with the input image as follows:

- Crop and resize the image to the fixed spatial extent of 512×512. The size of the produced ground truth heatmaps is 128×128.
- Return the pixel values to the domain $[0, 1]$ by dividing the pixel values of each image by 255.
- Online data augmentation.

Online data augmentation is the transformation of the input image directly when the image is fed into the model training. Some online data augmentations: random rotation $([-40°, 40°])$, random scale $([0.7, 1.3])$, random translation $([40, 40])$.

4.3. Parameters settings

In training process, the following parameters are used:

- Optimizer Algorithm: Adam
- Learning rate: $1 * 1e - 3$
- Optimizer strategy: We start training the model with $\mathbb{L}2$ loss initially, then with the focal $\mathbb{L}2$ loss until the performance rejects to increase.
- Batch size: 8
- Number of epoch: 60

- Input image's size: $512 \times 512 \times 3$
- Label heatmap'size: $128 \times 128 \times 50$

4.4. Evaluation metrics

The Object Keypoint Similarity (OKS) metric is used for evaluating the performance of keypoint detection models on the COCO dataset.

The OKS metric takes into account both the localization accuracy of the predicted keypoints and their semantic similarity to the ground truth keypoints. It computes a score between 0 and 1 that reflects how well the predicted keypoints match the ground truth keypoints for a given object instance. The higher the score, the more accurate the keypoint predictions are.

The use of OKS metric is important because it provides a standardized and quantitative way to evaluate the performance of different keypoint detection models on the same dataset. This allows researchers to compare the accuracy of different models and to track the progress of the field over time.

Moreover, the OKS metric is a key component in evaluating models for the COCO Object Detection Challenge, which helps to drive progress in the field by encouraging researchers to develop more accurate and effective keypoint detection models.

Average Precision (AP):	
AP	% AP at OKS$=.50:.05:.95$ (primary challenge metric)
AP^{50}	% AP at OKS$=.50$ (loose metric)
AP^{75}	% AP at OKS$=.75$ (strict metric)
AP Across Scales:	
AP^{M}	% AP for medium objects: $32^2 <$ area $< 96^2$
AP^{L}	% AP for large objects: area $> 96^2$

Where OKS is object keypoint similarity function, that performs the similar function as IoU. The formulation OKS:

$$OKS = \exp(-\frac{d_i^2}{2s^2k_i^2}) \tag{5}$$

- d_i is the euclidian distance between ground truth keypoint and predicted keypoint.
- s is scale: the square root of the object segment area.
- k is per-keypoint constant that controls fall off.

5. Results and discussion

5.1. Results on the MS-COCO dataset

In table 3, while our model did not achieve state-of-the-art (SOTA) performance like the bottom-up methods such as HigherHRNet [24], we obtained acceptable results and

Table 3. Results on the MS-COCO 2017 test-dev set

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
CMU-Pose	60.5	83.4	66.4	55.1	68.1
PersonLab	68.7	89.0	75.4	64.1	75.5
HigherHRNet	72.1	89.5	78.4	68.1	77.5
BK-Pose + $\mathbb{L}2$	57.3	78.2	64.7	54.3	65.7
BK-Pose + $\mathbb{FL}$	60.2	81.3	66.2	54.8	67.6

Table 4. Accuracy of localization of different keypoints on the MS-COCO 2017 test-dev set

Method	**Head**	**Shoulder**	**Elbow**	**Wrist**	**Hip**	**Knee**	**Ankle**	**Heel**	**Foot index**
BK-Pose + $\mathbb{L}2$	89.1	88.7	88.2	85.6	76.1	72.2	60.3	50.2	45.7
BK-Pose + $\mathbb{FL}$	91.8	90.1	89.8	87.4	79.5	75.4	63.0	51.6	46.2

found that it was suitable for application in classroom environments. The reasons lie in the nature of classroom activities: (1) The students default posture is sitting.
(2) Their type of activities is limited, compared to the variety of unusual activities (such as fighting or other sports). These facts significantly reduce the likelihood of false positives, thus improve accuracy and reliability of pose estimation. Table 4 presents the accuracy of localization of different keypoints on the MS-COCO 2017 test-dev. Due to the characteristic of BK-SAD dataset presented in subsection 4.1, our model obtains higher accuracy of the upper body detection, makes it particularly well-suited for classroom environments.

Based on table 3 and 4, we can conclude that the focal $\mathbb{L}2$ loss function outperforms $\mathbb{L}2$ loss function. Our results indicate that the focal $\mathbb{L}2$ loss function produced a higher accuracy rate compared to the $\mathbb{L}2$ loss function, suggesting that incorporating the focal term can enhance the model's ability to learn from hard examples. These findings suggest that the focal $\mathbb{L}2$ loss function can be an effective alternative to the L2 loss function for HPE tasks.

Table 5. Parameters and FPS against the SOTA bottom-up method on Jetson AGX Xavier [2]

Method	**#parameters**	**FPS**
PersonLab	68.7M	0.4
CMU-Pose	40.5M	0.2
HigherHRNet	28.6M	1.2
BK-Pose	**2.4M**	**11**

In this article, we conducted training exclusively on the MS-COCO dataset and BK-SAD dataset to build all our models from scratch. It has been observed that most state-of-the-art (SOTA) methods rely on the Hourglass backbone, which tends to have a large number of parameters. For example, while CMU-Pose [12] is well-known for

[2] https://developer.nvidia.com/embedded/jetson-agx-xavier-developer-kit

its ability to detect keypoints in real-time, it requires an NVIDIA GeForce GTX-1080 GPU to achieve a speed of 8.8 FPS.

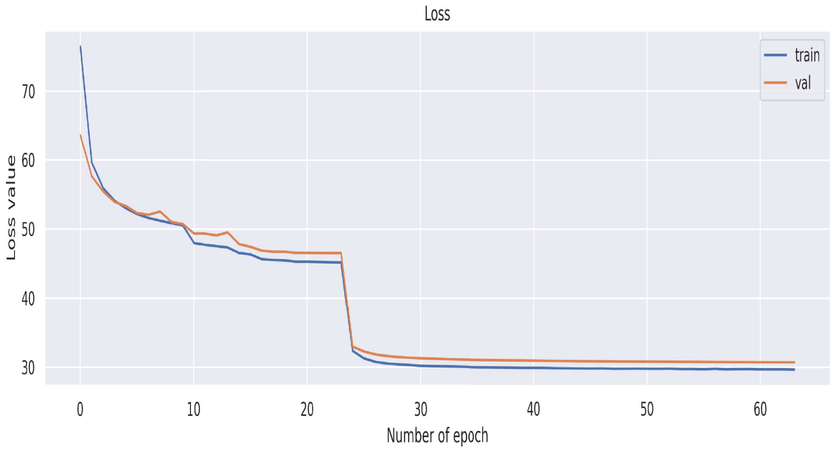

Fig. 6. Loss of BK-Pose model in 60 epochs on both training and validation set.

To evaluate the effectiveness of our proposed method, we compared the number of parameters and the frames per second (FPS) of some popular SOTA bottom-up models with an image size of 512×512 on the Jetson AGX Xavier device. Detailed comparison results are presented in table 5.

Our training process took a total of 5 days to complete. We initially trained the BK-Pose models using $\mathbb{L}2$ loss and subsequently fine-tuned them using the focal $\mathbb{L}2$ loss until there was no further improvement in performance. Due to the imbalance between the “easy” and “hard” keypoints, the training loss experienced slow convergence. Specifically, as depicted in Fig.6, the model’s loss exhibited a significant decrease at epoch 24 when $\mathbb{L}2$ loss was employed, signifying the successful acquisition of “easy” keypoints. Subsequently, we continued to train the model with focal $\mathbb{L}2$ loss until epoch 60 to allow the model to learn the “hard” keypoints.

5.2. Results on BK-SAD dataset

Based on table 6 and 7, we can conclude that BK-Pose model has a good performance on the BK-SAD dataset. This finding suggests that the incorporation of training with the MS-COCO dataset can be leveraged to enhance the accuracy of the model. Additionally, the comparative evaluation reveals that the focal $\mathbb{L}2$ loss function produced a higher accuracy rate compared to the $\mathbb{L}2$ loss function.

In table 8, we evaluate BK-SAD dataset by using ST-GCN [25] model in skeleton based action recognition experiments. We also experiment on two large-scale action recognition datasets: RGB+D [26], SBU Kinect Interaction [27]. The fact that the ST-GCN model can work well on BK-SAD datasets substantiates its effectiveness as a supplementary dataset for addressing the HPE problem.

Table 6. Results on the BK-SAD test set

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
BK-Pose + $\mathbb{L}2$	73.3	90.2	79.6	69.3	79.2
BK-Pose + $\mathbb{FL}$	74.2	91.6	80.5	70.4	80.7

Table 7. Accuracy of localization of different keypoints on the BK-SAD test set

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Heel	Foot index
BK-Pose + $\mathbb{L}2$	91.6	89.9	88.5	87.3	78.7	75.9	64.1	57.4	47.9
BK-Pose + $\mathbb{FL}$	92.7	90.8	90.2	89.3	80.7	77.2	66.4	58.8	49.2

Table 8. Activity recognition comparisons across 3 datasets via ST-GCN model

Dataset	Accuracy
SBU Kinect Interaction	52.8
NTU RGB+D	88.3
BK-SAD	**85.7**

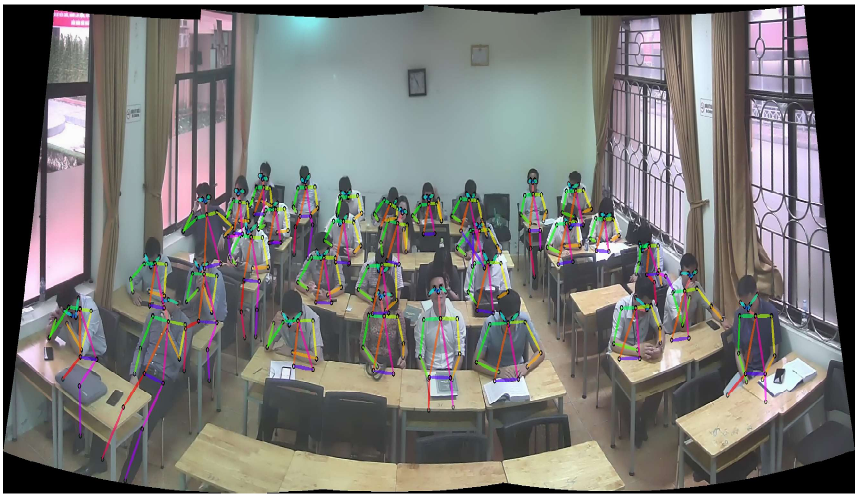

Fig. 7. Result at Academy of Politics Region I.

We conducted an extensive experiment to evaluate the efficacy of our proposed model at the Academy of Politics Region I, located at 15 Khuat Duy Tien, Thanh Xuan, Hanoi. Specifically, we captured keypoints of some classes consisting of 30 $\sim$ 40 students each. Our model was capable of detecting the upper part of the students' bodies and yielded promising results (Fig.7).

5.3. Failure cases

In spite of its overall performance, it still faces challenges that need to be addressed, including dense crowds, overlapping keypoints, and complex postures.

As demonstrated in Fig.8, the model is able to detect most keypoints, but they may be inaccurately located. Complex postures, such as dancing and walking, the model produces more false positive keypoints. The highest error rates occur in the legs and arms when they are hidden or in a crowded area.

These challenges can be attributed to two main factors: the accuracy of the BK-Pose model and the Keypoint Assignment Algorithm. We are currently working to address both of these issues to improve the model's performance.

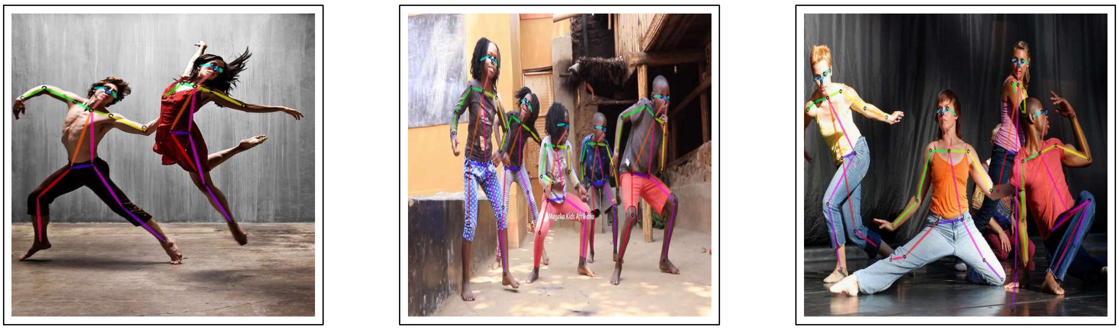

Fig. 8. Failure cases in "hard" keypoint of social activity.

6. Conclusions

In this study, we propose the BK-Pose bottom-up architecture for solving the multi-person pose estimation problem. The main contributions of our work are two-fold: (1) We introduce a lightweight bottom-up deep learning model that can be deployed on edge devices and can achieve high FPS rates while maintaining acceptable accuracy in suitable environments. (2) We incorporate a novel focal $\mathbb{L}2$ loss technique to balance between "hard" and "easy" keypoints samples during training. Our experimental results demonstrate that the proposed model has the acceptable accuracy comparing with existing models while being computationally efficient. Our approach can be applied in various environments, such as classroom or other public places for surveillance purposes and human activity recognition.

For future work, there are several promising directions to explore. With regard to the loss function, we only explore two loss functions in this study. Therefore, it is worthwhile to experiment with and compare other loss functions to identify better functions for this problem. In addition, with respect to the model architecture, we could replace the upscale module with deconvolution modules to generate high-quality and high-resolution feature maps and heatmaps, which may further enhance the model's performance.

Acknowledgement

We express our gratitude to the Ministry of Science and Technology for funding the national research project titled "Development of Digital Transformation model for Smart School" codename KC-4.0-06/19-25.

References

[1] Q. Dang, J. Yin, B. Wang, and W. Zheng, "Deep learning based 2D human pose estimation: A survey," *Tsinghua Science and Technology*, vol. 24, no. 6, pp. 663–676, 2019. doi: 10.26599/TST.2018.9010100

[2] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. doi: 10.1109/CVPR.2017.395 pp. 3711–3719.

[3] S. Huang, M. Gong, and D. Tao, "A coarse-fine network for keypoint localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. doi: 10.1109/ICCV.2017.329 pp. 3047–3056.

[4] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. doi: 10.1109/CVPR.2019.00584 pp. 5686–5696.

[5] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018. ISBN 978-3-030-01231-1 pp. 472–487.

[6] J. Wang, X. Long, Y. Gao, E. Ding, and S. Wen, "Graph-PCNN: Two stage human pose estimation with graph pose refinement," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020. ISBN 978-3-030-58621-8 pp. 492–508.

[7] Y. Cai, Z. Wang, Z. Luo, B. Yin, A. Du, H. Wang, X. Zhang, X. Zhou, E. Zhou, and J. Sun, "Learning delicate local representations for multi-person pose estimation," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 455–472.

[8] U. Iqbal and J. Gall, "Multi-person pose estimation with local Joint-to-Person associations," in *Computer Vision – ECCV 2016 Workshops*, G. Hua and H. Jégou, Eds. Cham: Springer International Publishing, 2016. ISBN 978-3-319-48881-3 pp. 627–642.

[9] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. doi: 10.1109/ICCV.2017.256 pp. 2353–2362.

[10] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. doi: 10.1109/CVPR.2016.533 pp. 4929–4937.

[11] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016. ISBN 978-3-319-46466-4 pp. 34–50.

[12] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis amp; Machine Intelligence*, vol. 43, no. 01, pp. 172–186, jan 2021. doi: 10.1109/TPAMI.2019.2929257

[13] D. Yu, K. Su, J. Sun, and C. Wang, "Multi-person pose estimation for pose tracking with enhanced cascaded pyramid network," in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019. ISBN 978-3-030-11012-3 pp. 221–226.

[14] S. Kreiss, L. Bertoni, and A. Alahi, "PifPaf: Composite fields for human pose estimation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. doi: 10.1109/CVPR.2019.01225 pp. 11 969–11 978.

[15] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy, "PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018. ISBN 978-3-030-01264-9 pp. 282–299.

[16] M. Kocabas, S. Karagoz, and E. Akbas, "MultiPoseNet: Fast multi-person pose estimation using pose residual network," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018. ISBN 978-3-030-01252-6 pp. 437–453.

[17] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. L. Zhu, F. Zhang, and M. Grundmann, "Blazepose: On-device real-time body pose tracking," *ArXiv*, vol. abs/2006.10204, 2020.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. doi: 10.1109/CVPR.2016.90 pp. 770–778.

[19] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-Scale Structure-Aware network for human pose estimation," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018. ISBN 978-3-030-01216-8 pp. 731–746.

[20] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," *Advances in Neural Information Processing Systems*, vol. 2, no. January, pp. 1799–1807, 2014.

[21] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2D pose estimation using part affinity fields,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. doi: 10.1109/CVPR.2017.143 pp. 1302–1310.

[22] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, “Cascaded pyramid network for multi-person pose estimation,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. doi: 10.1109/CVPR.2018.00742 pp. 7103–7112.

[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. doi: 10.1109/ICCV.2017.324 pp. 2999–3007.

[24] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, “HigherHRNet: Scale-Aware representation learning for Bottom-Up human pose estimation,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. doi: 10.1109/CVPR42600.2020.00543 pp. 5385–5394.

[25] W. Peng, J. Shi, and G. Zhao, “Spatial temporal graph deconvolutional network for Skeleton-Based human action recognition,” *IEEE Signal Processing Letters*, vol. 28, pp. 244–248, 2021. doi: 10.1109/LSP.2021.3049691

[26] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, “NTU RGB+D: A large scale dataset for 3D human activity analysis,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. doi: 10.1109/CVPR.2016.115 pp. 1010–1019.

[27] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, A. Natsev, M. Suleyman, and A. Zisserman, “The kinetics human action video dataset,” *ArXiv*, vol. abs/1705.06950, 2017.

Manuscript received 24-10-2022; Accepted 12-06-2023.

■

Van Giang Nguyen graduated from Hanoi University of Science and Technology in 2021. He is currently working toward a Master Degree of Data Science at the School of Information and Communication Technology, Hanoi University of Science and Technology. Research field: Computer Vision, Machine Learning. E-mail: giangnv.soict.hust@gmail.com

Chan Hung Nguyen received B. Eng and M.Eng degree in Electronics and Telecommunications at Hanoi University of Science and Technology, in 1995 and 1997, respectively. He obtained the degree of Ph.D. in Telematics from the University of Valladolid, Spain in 2002. He is now working at the Institute of Control Engineering and Automation – ICEA, Hanoi University of Science and Technology, Hanoi, Vietnam. He authored over 30 international publications on various ICT topics. Assoc. Prof. Hung was former member of IEEE ComSoc, IEEE Computer Society and oversea member of Japanese IEICE. E-mail: hungnc@gmail.com

Quang Dich Nguyen received a B.S. degree in electrical engineering from the Hanoi University of Technology in 1997. He received an M.S. degree in electrical engineering from the Dresden University of Technology, Dresden, Germany and a Ph.D from Ritsumeikan University, Kusatsu, Japan, in 2003 and 2010, respectively. Since 2000, he has been with the Hanoi University of Science and Technology, where he is currently an Associate Professor and Executive Dean of the Institute for Control Engineering and Automation. His research interests include magnetic bearings, self-bearing motors, and sensorless motor control.

Cong Dong Trinh received his Bachelor's degree in Electronics and Telecommunications Engineering from the Military Technical Academy in 2007. He obtained his master's degree in 2011. He graduated in Electronics and Telecommunications from Hanoi Open University. He is currently working at the Institute of Control and Automation Engineering - ICEA, Hanoi University of Science and Technology. Main field of study: Electronic engineering; artificial intelligence in smart agriculture; Industrial IoT applications; research on environmental monitoring systems. Email: dong.trinhcong@hust.edu.vn

BK-POSE: MỘT MÔ HÌNH NHẸ ƯỚC LƯỢNG DÁNG NGƯỜI

Nguyễn Văn Giảng, Nguyễn Chấn Hùng, Nguyễn Quang Địch, Trịnh Công Đồng

Tóm tắt

Ước lượng dáng người là một chủ đề cốt lõi trong nghiên cứu thị giác máy tính, có ứng dụng đa dạng trong nhiều lĩnh vực. Trong nghiên cứu này, chúng tôi giới thiệu một kiến trúc học sâu nhẹ mang tên BK-Pose, nhằm nhận dạng các khớp con người từ hình ảnh và video, và có thể triển khai trên các thiết bị tính toán biên. Công trình của chúng tôi góp phần quan trọng theo hai hướng: (1) Chúng tôi đề xuất một mô hình học sâu nhẹ, áp dụng kiến trúc từ dưới lên (bottom-up), thích hợp cho việc triển khai trên các thiết bị tính toán biên. Mô hình này đạt được tốc độ khung hình mỗi giây (FPS) cao, đồng thời vẫn đảm bảo độ chính xác ở mức chấp nhận được trong môi trường thích hợp. (2) Chúng tôi áp dụng một kỹ thuật hàm mất mát mới, gọi là focal L2 loss, để cân bằng giữa các mẫu khớp "khó" và "dễ" trong quá trình huấn luyện.

Chúng tôi đã tiến hành đánh giá hiệu năng của mô hình BK-Pose trong một môi trường lớp học để ước lượng các khớp của học sinh, và chứng minh tính hiệu quả của mô hình trong bài toán này. Kết quả nghiên cứu của chúng tôi có thể được sử dụng cho các nghiên cứu tiếp theo về nhận dạng hành vi sử dụng kiến trúc được đề xuất.

Từ khóa

Ước lượng dáng người, ước lượng dáng, học sâu, mạng nơ-ron tích chập.