JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 39, 1403-1420 (2023)
DOI: 10.6688/JISE.202311_39(6).0010

Multi-Person Pose Estimation using an Ordinal Depth-Guided Convolutional Neural Network

YI-YUAN CHEN[1], KUOCHEN WANG[1,2,+], HAO-WEI CHUNG[3,4], CHIEN-CHIH CHEN[5], BOHAU HUANG[1] AND I-WEI LU[2]

[1]*Department of Computer Science*
[3]*Department of Biological Science and Technology*
National Yang Ming Chiao Tung University
Hsinchu, 300 Taiwan
[2]*Center for Fundamental Science*
Kaohsiung Medical University
Kaohsiung, 807 Taiwan
[4]*Department of Pediatrics*
Kaohsiung Medical University Hospital
Kaohsiung, 807 Taiwan
[5]*Department of Industrial and Information Management*
National Cheng Kung University
Tainan, 701 Taiwan

Monocular 2D multi-person pose estimation in videos is essential for applications such as surveillance, action recognition, kinematics analysis, and medical diagnosis. Existing state-of-the-art offsets-based methods extract temporal features from offsets in consecutive predicted rough skeletons for better preciseness in fine-tuned the skeletons. However, the precision of existing single image-based models of rough skeleton prediction, such as HRNet, dropped due to shifting of target persons in propagated bounding boxes and resulted in inconsistent estimated poses in consecutive frames. To conquer this problem, we proposed an *Ordinal Depth-Guided-Convolutional Neural Network* (ODG-CNN) to address the issue. The proposed ordinal depth guides the Ordinal Depth-Guided Block (ODGB) in the ODG-CNN to reweight features for target persons in bounding boxes. Experiment results on the PoseTrack 2018 dataset indicate that the proposed ODG-CNN achieves the highest performance in terms of mean Average Precision (*mAP*). The proposed ODG-CNN is suited for applications, such as use of telehealth for early detection and intervention of developmental delays in children, which needs high accuracy of video-based estimated poses.

***Keywords*:** convolutional neural network (CNN), human pose estimation (HPE), multi-person pose estimation, ordinal depth, video-based HPE

1. INTRODUCTION

Human pose estimation (HPE) [1] has been studied in computer vision for years. It is a fundamental task for applications, such as surveillance [2], action recognition [3], and kinematics analysis [4], and medical diagnosis [5], which need precise skeletons across video frames. Most of the researches focusing on estimating human poses in a single image only considers the spatial context of images. While the precision of the image-based methods in static and high-quality images has improved considerably in recent years, perfor-

Received May 31, 2022; revised October 18, 2022; accepted January 7, 2023.
Communicated by Chun-Rong Huang.
[+] Corresponding author: kwang@cs.nctu.edu.tw

mance drops in video frames are inevitable due to blur, video defocus, and frequent pose occlusions [17]. In contrast, video-based HPE approaches take advantages of both spatial and temporal features for better performance. In comparison with single-person pose estimation, multi-person pose estimation is more challenging due to the occlusion problem between persons.

Current methods of HPE fall into two categories: bottom-up methods and top-down methods. Bottom-up methods, such as OpenPose [13], locate joints in an image followed by grouping joints into persons who they belong to. Top-down approaches iteratively estimate a pose with samples cropped by bounding boxes predicted by object detectors, such as Faster-RCNN [6] and YOLOv3 [7]. Generally, top-down approaches perform better than bottom-up approaches which trade-off time efficiency for joints precision. State-of-the-art video-based HPE approaches focus on extracting temporal features from image-based pose estimators. The LSTM-based method [14] applies long short-term memory (LSTM) learning and transferring temporal knowledge of a single person from the previous frames to the current frame. Offset-based methods [15-17] employ deformable convolution networks [22] learning pose offsets among adjacent frames for refining estimated poses of the current frames based on single-frame pose estimators as well.

1.1 Problem Statement

Although offset-based methods [15-17] have achieved the state-of-the-art precision of skeletons, these methods employ a bounding box propagation strategy for retrieving skeletons of a person among consecutive frames which trades-off consistent bounding boxes for precisions. The strategy uses the same bounding box as the one in the current frame for addressing the missing bounding boxes issue; however, people shifting within a propagated bounding box is inevitable and leads to target persons not locating at the center of the box. Such an occurrence is equivalent to the bounding box shifting issue [10] that a target person is not locating at the center of its predicted bounding box. The issue drops the performance of top-down approaches due to the approaches being trained with target persons located at the center of bounding boxes. Furthermore, the occurrence gets worse when crowded people are in bounding boxes.

1.2 Contribution

To overcome the bounding box shifting issue, we proposed a high-precision ordinal depth-guided-convolutional neural network (ODG-CNN) for multi-person pose estimation. Firstly, the proposed ODG-CNN is supervised with the proposed ordinal depth label. The proposed labeling algorithm defines the ordinal depth of the target person in a bounding box as front, middle, or back, relative to its nearest person when he/she exists. We also proposed an Ordinal Depth-Guided Block (ODGB). The predicted ordinal depth is forwarded to the proposed ODGB which adaptively reweights features spatial-wisely and channel-wisely according to the ordinal depth and input features.

In summary, the contributions of this paper are two-fold:

- We proposed ODG-CNN which addresses the issue of imperfect propagated bounding boxes in adjacent frames. Our approach achieves state-of-the-art results on the PoseTrack 2018 dataset [21].
- We proposed an ordinal depth labeling algorithm that defines an ordinal depth of a person

to its nearby persons in its bounding boxes. The ordinal depth is augmented information for the proposed ODGB to reweight features for better estimated poses.

1.3 Paper Outline

The rest of this paper is organized as follows. Section 2 presents the background and related work. In Section 3, we describe the details of the proposed method. Section 4 depicts implementation details of the experiments and experiment results. Finally, Section 5 gives conclusions and future work.

2. BACKGROUND AND RELATED WORK

In this section, we briefly introduce the self-attention mechanism (SAM) which was employed in our method. Then related work is reviewed.

2.1 The Self-Attention Mechanism

The self-attention mechanism (SAM) [31] learns correlations within inputs by transforming the inputs into query, key, and value. Attention scores are calculated through matrix multiplication of the key and the query. Self-attention is completed by adding the input data and the output of matrix multiplication between the attention scores and the value. W. Xiaolong *et al*. [24] proposed a non-local network (NLNet) sharing the idea of the self-attention mechanism for learning long-range dependencies point pairs in feature maps, as shown in Fig. 1 [25]. The value, key, and query are corresponding to the outputs of transformation functions, W_v, W_k, W_q. After transformation, feature maps are reshaped for matrix multiplications and are shown as feature dimensions. For example, $C \times H \times W$ denotes a feature map with channel number C, height H, and width W. Finally, the self-attention mechanism is completed after the addition of the multiplicative features (obtained from the convolution of $C \times H \times W$ and W_z, a weight matrix) and input features.

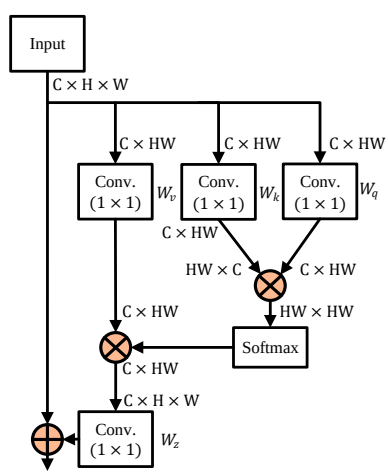

Fig. 1. Architecture of an extension of the self-attention mechanism, a non-local block. Conv. denotes 2D convolution, ⊕ denotes broadcast element-wise addition, and ⊗ denotes matrix multiplication.

2.2 Related Work

Fig. 2 shows classification of related studies on video-based HPE. In video-based HPE, the recurrent neural network (RNN) and the convolutional neural network (CNN) are used to extract temporal features. RNN-based methods connect nodes that form a directed graph along a temporal sequence. CNN-based methods learn temporal features in video sequences by concatenating features in consecutive frames. A kind of RNN, LSTM is employed for memorizing long short-term temporal features. In CNN, four types of methods are applied to handle temporal features: optical flow, 3D CNN, propagation and pose offset. The optical flow-based method makes use of a technique tracking motion of the image pixel with pixel intensity. The 3D-CNN-based method lets networks learn temporal features by themselves. The propagation-based method propagates information across frames as temporal guidance [16]. The method of pose offset learning temporal features are based on consecutive probability density maps of joints, such as heatmaps. Specifically, most pose offset-based methods are based on consecutive rough poses estimated from off-the-shelf single image-based pose estimation methods independently, which will make these methods not robust to imperfect propagated bounding boxes. Although the proposed ODG-CNN method is also a pose offset-based method, it takes advantage of the ordinal depth information which will make the proposed method more robust to imperfect propagated bounding boxes.

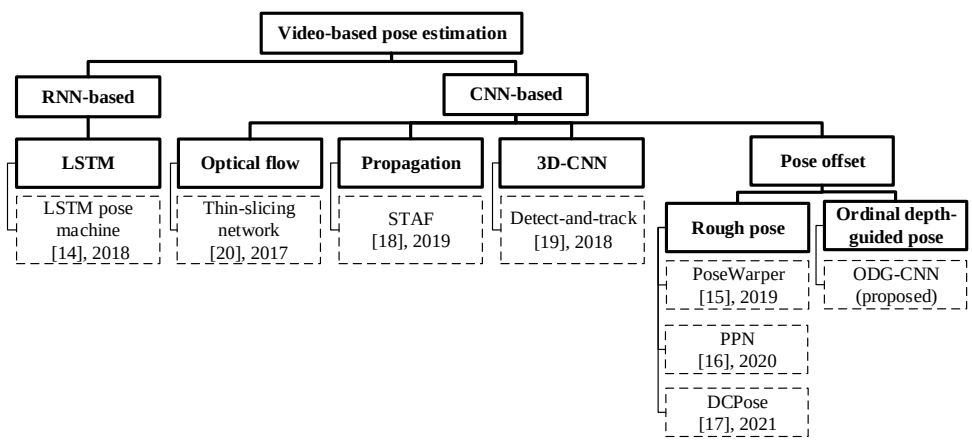

Fig. 2. Classification of related studies on video-based HPE.

2.2.1 Bottom-up HPE

Propagation-based: Bottom-up based HPE methods detect joints for all persons at once in an image without identifying instances and then group the predicted joints into persons who they belong to [13]. This enables bottom-up methods to achieve real-time pose estimations. To further extract temporal features, Y. Raaj *et al.* [18] proposed a temporal affinity field (TAF) based on the partial affinity field (PAF) proposed by OpenPose [13]. TAFs are dependent on PAFs and extracted features from the current and the previous frames, as well as TAFs from the previous frame [18].

2.2.2 Top-down HPE

3D-CNN-based: R. Girdhar *et al.* [19] proposed a 3D human pose predictor by extending Mask R-CNN [32]. The operations of 2D convolution in the model are transformed to 3D convolution for extracting temporal features in a short clip. 3D filters are initialized by extending the parameters of a pretrained 2D network with a center initialization method. The proposal network proposes tube candidates with instance-specific features for regression (bounding box proposal), classification (human or not), and pose estimation.

LSTM-based: Y. Luo *et al.* [14] proposed a method extracting temporal features in 2D single-person videos. The approach extends CPM [9] for learning estimated heatmaps. Gaussian masks focusing on each joint are added with previous results and the estimated joint heatmaps of the current frame are fed to LSTM for learning dynamics and location of each joint. The model achieves high precision in the single-person video; however, extracting temporal features of target persons in a multi-person environment is severely impacted by occlusion [17].

Optical flow-based: J. Song *et al.* [20] warps heatmaps estimated by CPM [9] as a rough estimation. Then, aligned heatmaps are generated by warping the heatmaps of the same persons in adjacent frames to the current frames with dense optical flow, pixel-wise flow vectors, for the stability of joint predictions. In the spatio-temporal inference stage, it formulates a spring energy model describing the action ranges of each joint defining a human skeleton. With multiple warped poses inferred from the previous and the subsequent frames via optical flow, the proposed model determines the final joints locations based on the belief propagation algorithms [34].

Pose offset-based: G. Bertasius *et al.* [15] predicts auxiliary poses by employing a strategy of propagating a bounding box of a current frame to its adjacent frames which reduces chances of missing bounding boxes from human detectors. With an auxiliary pose and a pose in previous and keyframes, its dilated deformable convolutional module learns multi-scale offset between the poses for fine-tuning the pose of the keyframe. Y. Liu and J. Chen [16] proposed a light-weighted model which contains pose propagation units (PPU) built upon dilated deformable convolutional modules [15]. The approach recurrently estimates poses in every short clip. Starting with the first frame, the final pose in one frame is propagated to an inference of the next frame (the next iteration) as guidance for pose refinement in the next frame. The elementwise addition of propagated heatmaps and the preliminary pose drives the dilated deformable convolutional module [15] to learn the weights between propagated and preliminary heatmaps in the training phase. Zhenguang Liu *et al.* [17] proposed a dual consecutive network making use of auxiliary poses from previous and next frames with deformable convolutional module which augments pose information from the next frame for adjustment and won the first place in multi-person challenges, PoseTrack 2017 [21].

Pose offset-based methods currently perform better than the other methods in the multi-person dataset, PoseTrack [20]. It typically employs image-based pose estimators as the backbone and fine-tunes the rough poses from it. These methods employ a bounding box propagation strategy under the assumption that a person in consecutive frames changes

slightly and pose estimators produce a pose to the same person in the consecutive frames with the propagated bounding box. However, when it comes to a target person and multiple persons appearing in a propagated bounding box of the target person, the target person shifting in the bounding box due to either the target person's self-movement or camera motion in consecutive frames may cause top-down pose estimators unstably to predict the pose in the propagated bounding box. Furthermore, the erroneous skeleton will cause the offset-based methods to fail to refine joints with temporal features from erroneous ones.

In summary, the main differences between state-of-the-art studies and the proposed ODG-CNN method in monocular video-based HPE are shown in Table 1. In this work, we proposed an ordinal depth-guided-convolutional neural network (ODG-CNN) which explores the ordinal depth of a target person in the multi-person image for further feature reweighting. The proposed ODG-CNN model estimates the poses in the bounding box by referring to the predicted ordinal depth information that enhances the precision in propagated bounding boxes for better refinement of poses.

Table 1. Comparison of state-of-the-art studies in monocular video-based HPE.

Method	Temporal resource	Pros	Cons
Detect-and-Track [19], 2018	3D Mask R-CNN	Bidirectional temporal features for both detection and estimation	Lower precision
STAF [18], 2019	Temporal affinity field (TAF) [18] and recurrent CNN	Lower computations in the efficient version (26 fps)	More computation in the high precision version
PoseWarper [15], 2019	Consecutive heatmaps	Single directional temporal fine-tuning	Not robust to propagated bounding boxes
DCPose [17], 2021	Consecutive heatmaps	Bidirectional temporal finetuning	Not robust to propagated bounding boxes
ODG-CNN (proposed)	Ordinal depth-guided consecutive heatmaps	– Bidirectional temporal fine-tuning – Ordinal depth-guided for feature reweighting	More computation

3. PROPOSED ORDINAL DEPTH-GUIDED CONVOLUTIONAL NEURAL NETWORK

In this section, we first introduce the proposed ordinal depth labeling algorithm which is to differentiate person-specific ordinal depths between people in an image. Then, the proposed ordinal depth-guided block (ODGB) and the details of the proposed ODG-CNN are described.

3.1 Ordinal Depth Labeling Algorithm

Although the model of R. Khirodkar *et al.* [12] improves the performance of HRNet [11] with multi-hypothesis, the condition λ does not contain additional information but 0 and 1. Therefore, inspired by the idea of ordinal relations in G. Pavlakos *et al.* [27] that annotates ordinal relations between joint pairs in 2D images, we proposed an algorithm to label the ordinal depth of a person to its nearby persons with one-hot encoding. Due to the dataset such as PoseTrack 2018 does not has annotations of ordinal depth as the manual annotations in G. Pavlakos *et al.* [27], we provided an ordinal depth labeling algorithm and defined the ordinal depth of each person according to the number of visible joints and the

head size. The ordinal depth is defined as *front*, *middle*, or *back*, which is a relative depth to its nearest person when he/she exists. In other words, the algorithm defines a person's ordinal depth among crowded people by his/her distance to the camera.

In a given image $I \in R^{H\times W\times 3}$ composed of N poses of persons $P = \{p_i\}$ and ground-truth bounding box $B = \{b_i\}$, where $i \in \{0, 1, \ldots, N-1\}$, the proposed model estimates the ordinal depth of a target person p_i from an image cropped with the corresponding bounding box b_i. Afterwards, the ordinal depth is forwarded to the ordinal depth-guided block (ODGB) for feature extraction and pose estimation. Let take a pose of a person as a target person, $p_T \in P$, and the corresponding bounding box $b_T \in B$, for example. We follow the method of R. Khirodkar *et al.* [12] defining nearby persons if a person has at least three visible joints locating within b_T and get a list of candidate C persons near to the p_T, where $C \le N$. The set of poses, $P_C = \{p_{ci}\}$, and the bounding boxes, $B_C = \{b_{ci}\}$, contain poses and bounding boxes of each candidate nearby the person, where $1 \le i \le C$. The distance of center coordinates between b_T and b_{ci}, $D(b_T, b_{ci})$, is applied to define the order of a nearby person from near to far. The descriptions can be formulated as Eq. (1) below,

$$i^* = \underset{i}{argmin}\, D(b_T, b_{ci}). \tag{1}$$

The candidate with the shortest distance, p_{i^*}, is defined as the nearest person p_N.

Following the selected nearest nearby person, we define the ordinal depth d_T of the target person by comparing the numbers of visible joints of p_T and p_N, denoted as v_T and v_N, respectively. As shown in Eq. (2), we label d_T as "front" if the difference between v_T and v_N is greater than T_j. If the difference between v_T and v_N is less than $-T_j$, d_T is labeled as "back." Otherwise, d_T is labeled as "middle," as shown in Eq. (2),

$$d_T = \begin{cases} front, & \text{if } v_T - v_N > T_j \\ back, & \text{if } v_T - v_N < -T_j \\ middle, & \text{else} \end{cases} \tag{2}$$

where T_j is the threshold of the number of visible joints and is set to 2 according to experiment data.

However, strictly following Eq. (2) may wrongly label ordinal depths. For instance, when a target person p_T is at the ordinal depth of 'front' relative to its nearest person p_N, p_T might be at the ordinal depth of 'back' when it is the nearest person to the other persons. Therefore, the head bounding box is also considered if the ordinal depth of a person is 'middle' according to Eq. (2). We compare the head sizes of the target person h_T and its nearest person h_N, and follow Eq. (3),

$$d_T = \begin{cases} front, & \text{if } h_T / h_N \ge T_{hf} \\ back, & \text{if } h_T / h_N \le T_{hb} \\ middle, & \text{else} \end{cases} \tag{3}$$

where T_{hf} and T_{hb} are the thresholds of the head sizes for defining ordinal depth "front" and "back," and are set to 1.69 and 0.59, respectively, according to experiment data. Note that the head size is defined as an area of a head bounding box by multiplying its width and length.

Fig. 3 is an example to illustrate the above situation and the usages of Eqs. (2) and (3) wing to the occlusion problem among three persons using the PoseTrack 2018 dataset. The green person (the person with a green head bounding box) is the target person p_T, the blue person is the nearest person p_{N1}, and the purple person is the second nearest person p_{N2}, where the number of visible joints $v_T = 10$, $v_{N1} = 9$ and $v_{N2} = 13$. The head sizes are $h_T = 74 \times 80$, $h_{N1} = 45 \times 51$, and $h_{N2} = 123 \times 121$. For the relationship between p_T and p_{N1}, according to Eq. (2) since $(v_T - v_{N1} = 1) < (T_j = 2)$, d_T would be labeled as "middle." In this situation, the head sizes of the target person (h_T) and the blue person (h_{N1}) should be compared. According to Eq. (3), since $(h_T/h_{N1} = 2.58) \geq 1.69$ ($T_{hf} = 1.69$), d_T would be labeled as "front." For the relationship between p_T and p_{N2}, according to Eq. (2), since $(v_T - v_{N2} = -3) < (-T_j = -2)$, d_T would be labeled as "back." Since d_T is labeled as "front" and "back" relative to p_{N1} and p_{N2}, respectively, d_T should be labeled as "middle" among the three persons. That is, the ordinal depths of the purple person p_{N2}, the green person p_T (the target person) and the blue person p_{N1} are "front", "middle", and "back", respectively, which correctly identifies the relationship among the three persons, as shown in Fig. 3.

Fig. 3. An example to illustrate ordinal depth labeling using the PoseTrack 2018 dataset.

3.2 Ordinal Depth-Guided Block

R. Khirodkar *et al.* [12] introduced a multi-hypothesis block (MHB) with condition input λ, where $\lambda = \{0, 1\}$, to deal with the number of limited bounding boxes due to low confidence caused by multiple persons. The model can produce primary hypothesis and residual hypothesis for target persons and its nearest person with $\lambda = 0$ and 1, respectively, from the same bounding box. However, the primary hypothesis requires target persons located at the center of bounding boxes. To improve R. Khirodkar *et al.* [12], we proposed an Ordinal Depth-Guided Block (ODGB), as shown in Fig. 4. The ODGB is a revised version of the global context block (GCB) proposed by Y. Cao *et al.* [25], an extension of the self-attention mechanism that learns long-range dependencies point pairs in feature maps but with a lighter structure than the non-local block [24]. Different from the GCB [25], the proposed ordinal depth is broadcast and cascaded with input features. The cascaded features are transformed as the key for computing attention scores with the original features. Finally, the output of the computation is decoded and added to the original features.

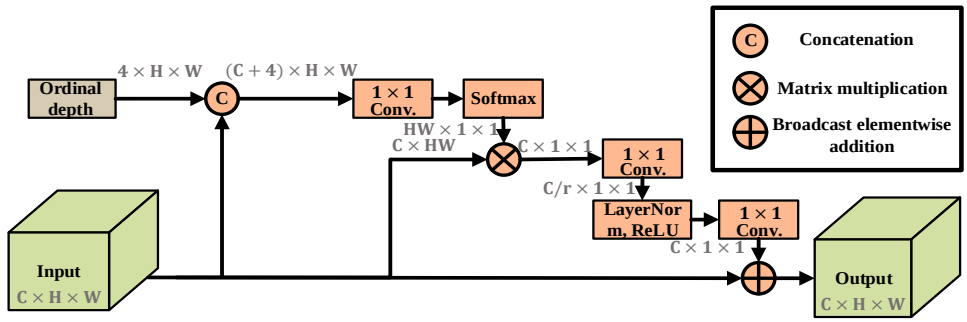

Fig. 4. The details of the proposed Ordinal Depth-Guided Block (ODGB).

The overview of the proposed ODG-CNN is shown in Fig. 4. The ordinal depth in the current frame is propagated along with bounding box propagation. Note that the ordinal depth is predicted from the feature map that contains the information of a person, which originates from the associated image. A more accurate result is forward to the offset-based fine-tuning network, as DCPose [17]. The framework of the proposed ordinal depth-guided-pose network (ODG-PNet) is shown in Fig. 6, where ODGBs are added to the last two stages of HRNet-48 [11]. As shown in Fig. 6, stage 1 (the 1st stage) extracts features from an image under different scales and then generates a lower resolution subnetwork with the resolution of the feature map decreased to a half size from 72×96 to 36×48 and the number of channels increased by two times from 48 to 96. Stage 2 (the 2nd stage) performs multi-scale fusion which aggregates two branches under different scales to output two fused feature maps under different feature sizes. In addition, it generates a lower resolution subnetwork with the resolution of the feature map decreased to a half size from 36×48 to 18×24 and the number of channels increased by two times from 96 to 192. Stages 3 and 4 also perform multi-scale fusion, similarly as stage 2. In stage 4, before it aggregates four branches under different scales, each branch passes through an ODGB to reweight the features via the predicted ordinal depth, and then it outputs one fused feature map. In summary, from Stages 1 to 3, the resolution of the newly generated subnetwork's feature map in each stage is gradually decreased to a half size ($72 \times 96 \rightarrow 36 \times 48 \rightarrow 18 \times 24 \rightarrow 9 \times 12$) and the number of channels is increased by two times ($48 \rightarrow 96 \rightarrow 192 \rightarrow 384$).

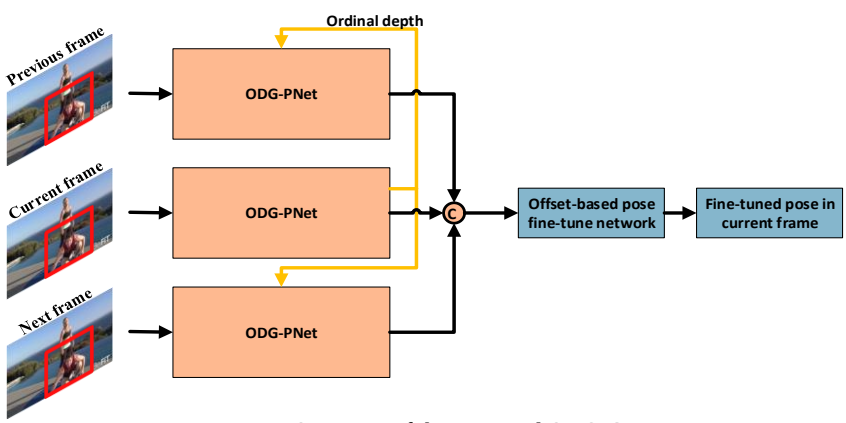

Fig. 5. Overview of the proposed ODG-CNN.

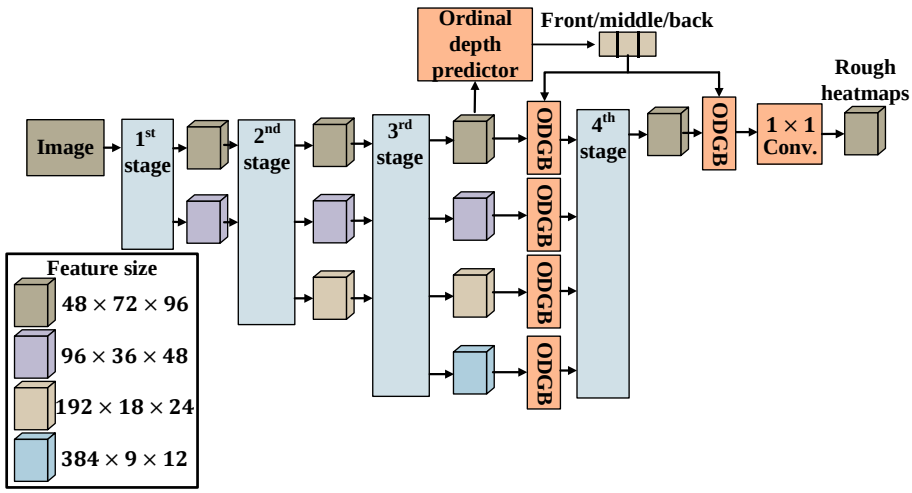

Fig. 6. The framework of the proposed ODG-PNet.

3.3 Implementation Details

3.3.1 Data augmentation

During training, we followed the data augmentation steps as those of HRNet [11]. Firstly, we randomly cropped targets persons into half bodies, then randomly flipped the image for learning symmetric limbs. After that, boxes were randomly rotated with degrees within [−45, 45], scaled with degrees within [−35%, 35%]. Finally, all boxes were resized to a fixed resolution of 288 for height and 384 for width.

3.3.2 Loss function

We employed the L2 loss function to calculate losses between predictions and ground-truth heatmaps pixel-wisely for all visible joints, which is a common pose estimation loss function as that in R. Khirodkar *et al.* [12]. Invisible joints are not involved in backward propagation. The loss function is defined as shown in Eq. (4),

$$L_H = \frac{1}{N} \times \sum_{j=1}^{N} v_j \times \| H_G(j) - H_p(j) \|^2 \tag{4}$$

where $H_G(j)$, $H_P(j)$ and v_j denote ground-truth heatmap, predicted heatmap, and an indicator for visibility of joint j. The number of joints (N) in the PoseTrack 2018 dataset is 15. 2D Gaussian with a standard deviation of 2 centering at the location of each ground-truth joint is applied for ground-truth heatmaps.

To supervise the ordinal depth of persons, we employed a focal loss function L_D from Tsung-Yi *et al.* [26] to learn classification tasks of the ordinal depth. The function is defined as shown in Eq. (5),

$$L_D = G_c \times -\alpha_c (1 - p_c)^{\gamma} \log(p_c) \tag{5}$$

where G_c denotes an indicator, which is 1 when the ground-truth label is class c. α_c is used to balance the number of samples in each class c. γ is a factor that allows the model to focus more on harder classification classes. p_c denotes the predicted probability of class c.

4. EXPERIMENT RESULTS AND DISCUSSION

In this section, we firstly describe the PoseTrack 2018 dataset [21] and evaluation metrics used. Then, experiment results and discussion are presented.

4.1 Dataset and Tools

The PoseTrack 2018 dataset [21] is a large-scale multiple-person in-the-wild for HPE and tracking in videos. The dataset has complicated movements of highly occluded people and camera motions. It is composed of 1,138 clips, including 153,615 annotations of the poses. The dataset was split into training, validation, and testing sets including 593, 170, and 375 clips, respectively. For the clips in the training split, 30 frames are densely annotated in the center of the clips. For the clips in the validation split, poses are annotated in every 4 frames. In addition, the annotations in training and validation datasets include human bounding boxes, head bounding boxes, locations of 15 joints of a person, and joint visibility of each joint.

4.2 Training Details and Setup

In training, we firstly loaded parameters of HRNet [11] pretrained on the COCO dataset [28] before training the proposed Ordinal Depth-Guided-CNN (ODG-CNN) model with PoseTrack 2018. Then, we preprocessed data with data augmentation described in Section 3.3.1. ODG-PNet was trained with the preprocessed data and the loss function described in Section 3.3.2. To evaluate the improvement of the ordinal depth block (ODGB) in both propagated and non-propagated bounding boxes, we compared the HRNet [11], MHPNet [12], and the proposed ODG-PNet model. To evaluate the improvement of temporal features, we also trained the proposed ODG-CNN by following the training details of ODG-PNet. We employed the Adam optimizer [29] with a learning rate starting with 0.0001 and decaying by 10% every 2 epochs. We trained each model for a batch size of 20 for 20 epochs. Each model was implemented with PyTorch [30] and trained with 1 Nvidia Tesla V100S-PCIE-32GB.

4.3 Evaluation Metrics

Average precision (AP) is used to evaluate the precision of each joint part in the PoseTrack benchmark [21]. With multiple body pose predictions, only the single pose with the highest percentage of correct keypoints (PCK) can be assigned to the ground-truth pose. The other predictions are considered as false positives. In detail, $\text{PCKh}_{0.5}$ was introduced in the PoseTrack 2018 benchmark for filtering out predicted poses that have values normalized by head length higher than 0.5. Best predicted poses out of the filtered poses are selected to match the corresponding ground-truth poses. The head length corresponds to 60% of the diagonal length of the ground-truth head bounding box [33]. If the distance between a predicted joint and the corresponding ground-truth joint is within 50% of the

head length, the prediction is considered correct, otherwise, incorrect. Finally, mean AP (*mAP*) is the mean of APs over 15 joints in PoseTrack 2018.

4.4 Evaluation Results and Discussion

4.4.1 Comparison with state-of-the-art approaches

The single image-based and video-based HPE's were evaluated on the validation set of the PoseTrack 2018 dataset, as shown in the upper part of Table 2. The experiment results of HRNet [11] and MHPNet [12] on the Posetrack 2018 validation dataset were trained by following their training details. As shown in Table 2, the proposed ODG-PNet achieved the highest performance of APs over all joints. That is, ODG-PNet achieved the highest *mAP* of 80.3%.

Comparison of the state-of-the-arts video-based HPE approaches and the proposed ODG-CNN are shown in the bottom part of Table 2. The reproduced performance results with the source code released from Z. Liu *et al.* [17] are also presented and marked with a superscript star. We reproduced DCPose [17] with the fine-tuned HRNet [11], as shown in the upper part of Table 2. ODG-CNN obtained *mAP* of 81.1%, which is higher than the performance of DCPose. In addition, *mAP* of the proposed ODG-CNN model is also higher than the reproduced performance of DCPose [17] with the same fine-tune network.

Table 2. Performance (*mAP* (%)) evaluation of single-image based and video-based state-of-the-arts models using the PoseTrack 2018 dataset.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean (*mAP*)
Single image-based multi-person pose estimation								
HRNet [11], 2019	83.4	86.5	82.2	77.1	76.8	77.8	72.8	**79.7**
MHPNet [12], 2021	83.4	86.5	82.0	76.3	78.4	78.1	72.3	**79.8**
ODG-PNet (proposed)	83.4	86.6	82.7	77.1	79.3	78.7	73.2	**80.3**
Video-based multi-person pose estimation								
STAF [18], 2019	NA	NA	NA	64.7	NA	NA	62.0	**70.4**
Pose Warper [15], 2019	79.9	86.3	82.4	77.5	79.8	78.8	73.2	**79.7**
DCPose [17], 2021	84.0 84.2*	86.6 86.8*	82.7 82.8*	78.0 77.6*	80.4 78.7*	79.3 78.3*	73.8 73.3*	**80.9** **80.5***
ODG-CNN (proposed)	83.9	87.0	82.9	77.8	80.6	79.7	74.4	**81.1**

* Denotes the reproduced performance results of DCPose [17] with the latest source code version released by the authors.

4.4.2 Evaluation of poses in propagated bounding boxes

Due to no relevant discussion on the performance of predicted poses in propagated bounding boxes, we had experiments of poses estimation in propagated bounding boxes with HRNet [11] and MHPNet [12], according to their descriptions and the proposed

ODG-PNet for the comparison of *mAP*, as shown in Table 3. In comparison with HRNet [11] and MHPNet [12], ODG-PNet achieved the highest *mAP* in propagated bounding boxes with a propagation distance of 1. In terms of computation complexity, the number of parameters of ODG-PNet is slightly higher than the other two approaches because of the ordinal depth predictor and the ODGBs used in ODG-PNet. In terms of multiply and accumulate operations (MACs), the number of MACs in ODG-PNet is also slightly higher than the other two approaches. However, the proposed ODG-PNet achieved the highest *mAP* in propagated bounding boxes.

Table 3. Performance (*mAP* (%)) evaluation of HRNet, MHPNet, and the proposed ODG-PNet in propagated bounding boxes.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean (*mAP*)	#Params	MACs
HRNet [11], 2019	80.4	83.5	79.2	74.1	74.1	74.9	69.9	**76.8**	63.6M	35.4G
MHPNet [12], 2021	80.3	83.6	79.2	73.5	75.7	75.3	69.5	**77.0**	63.7M	35.4G
ODG-PNet (proposed)	80.6	83.9	80.0	74.6	76.8	76.1	70.6	**77.7**	63.8M	35.8G

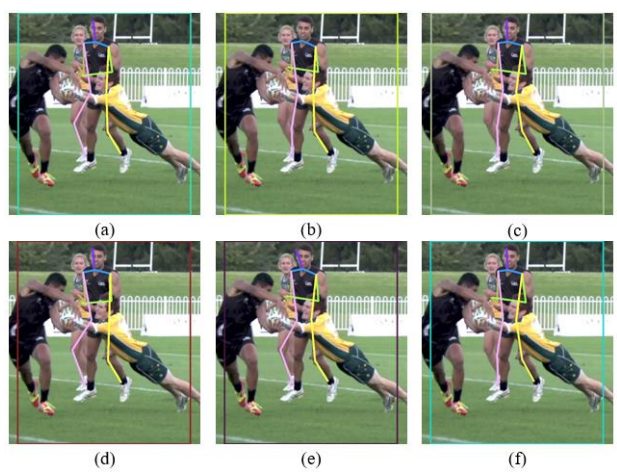

Fig. 7. Visualized pose estimation results of HRNet [11] (a), MHPNet [12] (b) and the proposed ODG-PNet (c) in the current bounding box, and HRNet [11] (d), MHPNet [12] (e) and the proposed ODG-PNet (f) in the propagated bounding box using the PoseTrack 2018 dataset.

We use an example to illustrate and to visualize the effectiveness of the proposed ODG-PNet. Fig. 7 shows visualized pose estimation results of HRNet [11] (a), MHPNet [12] (b)) and the proposed ODG-PNet (c) in the current bounding box, and HRNet [11] (d), MHPNet [12] (e) and the proposed ODG-PNet (f) in the propagated bounding box using the PoseTrack 2018 dataset. We found that in the current bounding box, the proposed ODG-PNet (c) and MHPNet [12] (b) have similar precision performance of pose estimation results, while HRNet [11] (a) incorrectly predicted the right knee (in pink) joint of the person in the back as that of the target person (the person in the middle). In addition, in the

propagated bounding box, the proposed ODG-PNet (f) has better pose estimation results, especially in the right knee joint (in pink) prediction of the target person (the person in the middle), compared to HRNet [11] (d) and MHPNet [12] (e). This is because in the propagated bounding box, the proposed ODG-PNet has additional "ordinal depth" information that can help predict the right knee joint of the target person, while the other two methods, HRNet [11] and MHPNet [12], both incorrectly predicted the right knee joint of the person in the back as that of the target person. This example illustrates and visualizes the effectiveness of the proposed ODG-PNet, compared to the other two methods.

4.4.3 Ablation studies

Ablation studies of ODG-PNet performance on the PoseTrack 2018 are shown in Table 4. In the first part, we studied combinations of two parameters, head size and the number of visible joints (#joints), in the ordinal depth algorithm. The results show that ODG-PNet trained with the ordinal depth defined with either parameter of head size or #joints is not higher than the combination of the two parameters. Furthermore, the combination in the order of #joints before head size has the best performance of APs overall. To validate the proposed ODGB, we also trained HRNet [11] employed with GCB [25], named as ODG-PNet w/o ODGB. The results show that the proposed ODG-PNet with ODGB improved *mAP* by 0.6%.

Table 4. Ablation studies of ODG-PNet performance (*mAP*(%)) on PoseTrack 2018.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean (*mAP*)
Ordinal depth algorithm								
ODG-PNet (head size only)	83.3	86.6	82.6	77.2	79.3	78.5	73.0	**80.3**
ODG-PNet (#joints only)	83.8	86.6	82.5	77.2	79.2	78.5	72.8	**80.3**
ODG-PNet (head size + #joints)	83.3	86.5	82.6	77.0	79.3	78.5	73.0	**80.3**
ODG-PNet (#joints + head size) (proposed)	83.4	86.6	82.7	77.1	79.3	78.7	73.2	**80.3**
Ordinal depth-guided block								
ODG-PNet w/o ODGB	82.9	86.3	82.1	76.3	78.5	78.4	72.3	**79.7**

5. CONCLUSION AND FUTURE WORK

5.1 Conclusion

This paper presents the proposed ordinal depth-guided-CNN (ODG-CNN) which provides a propagated bounding box with additional information of the ordinal depth for video-based multi-person pose estimation. Experiment results have shown that the proposed ODG-CNN achieves the higher *mAP* of 81.1% on the PoseTrack 2018 dataset than existing video-based multi-person pose estimation methods. In addition, experiment results have also shown that the proposed ODG-PNet has better performance in terms of *mAP* than existing single image-based multi-person pose estimation methods, such as

HRNet [11] and MHPNet [12] in propagated/non-propagated bounding boxes. This shows the robustness of the proposed ODG-PNet to estimating poses in propagated bounding boxes.

5.2 Future Work

In the proposed ODG-CNN, we simply reweight features for the target person by concatenating the proposed ordinal depth. In the future, we will further investigate tasks of a combination of computer vision (CV) and natural language processing (NLP), such as visual question answering (VQA) to answer sentences by fusing features of questions and images, for exploiting more spatial information with the proposed ordinal depth. In addition, we plan to implement the proposed ODG-CNN with datasets of real-world applications that need high *mAP* of video-based estimated poses, such as use of telehealth for early detection and intervention of developmental delays in children.

ACKNOWLEDGEMENTS

The supports by the Kaohsiung Medical University Hospital under Grants KMUH-SI10910 and by the Ministry of Science of Technology (MOST), Taiwan, under Grants MOST 110-2221-E-037-006 are greatly appreciated.

REFERENCES

1. C. Zheng *et al.*, "Deep learning-based human pose estimation: A survey," https://arxiv.org/abs/2012.13392, 2020.
2. S. Das, S. Sharma, R. Dai, F. Bremond, and M. Thonnat, "VPN: Learning video-pose embedding for activities of daily living," in *Proceedings of European Conference on Computer Vision*, 2020, pp. 72-90.
3. S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of AAAI*, 2018, pp. 1-9.
4. J. Xu, Z. Yu, B. Ni, J. Yang, X. Yang, and W. Zhang, "Deep kinematics analysis for monocular 3d human pose estimation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 899-908.
5. C. Chambers *et al.*, "Computer vision to automatically assess infant neuromotor risk," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 28, 2020, pp. 2431-2442.
6. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015, pp. 91-99.
7. J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," https://ariv.org/abs/1804.02767, 2018.
8. A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proceedings of European Conference on Computer Vision*, 2016, pp. 2597-2602.
9. S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines,"

in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724-4732.
10. H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 2334-2343.
11. K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693-5703.
12. R. Khirodkar, V. Chari, A. Agrawal, and A. Tyagi, "Multi-hypothesis pose networks: Rethinking top-down pose estimation," https://arxiv.org/abs/2101.11223, 2021.
13. Z. Cao, G. Hidalgo, T. Simon, S. -E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43, 2021, pp. 172-186.
14. Y. Luo *et al.*, "LSTM pose machines," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5207-5215.
15. G. Bertasius, C. Feichtenhofer, D. Tran, J. Shi, and L. Torresani, "Learning temporal pose estimation from sparsely-labeled videos," in *Advances in Neural Information Processing Systems*, 2019, pp. 3027-3038.
16. Y. Liu and J. Chen, "PosePropagationNet: Towards accurate and efficient pose estimation in videos," in *IEEE Access*, Vol. 8, 2020, pp. 100661-100669.
17. Z. Liu, H. Chen, R. Feng, S. Wu, S. Ji, B. Yang, and X. Wang, "Deep dual consecutive network for human pose estimation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 525-534.
18. Y. Raaj, H. Idrees, G. Hidalgo, and Y. Sheikh, "Efficient online multi-person 2D pose tracking with recurrent spatio-temporal affinity fields," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4615-4623.
19. R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran, "Detect-and-track: Efficient pose estimation in videos," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 350-359.
20. J. Song, L. Wang, L. Van Gool, and O. Hilliges, "Thin-slicing network: A deep structured model for pose estimation in videos," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5563-5572.
21. M. Andriluka *et al.*, "PoseTrack: A benchmark for human pose estimation and tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5167-5176.
22. J. Dai *et al.*, "Deformable convolutional networks," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 764-773.
23. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132-7141.
24. W. Xiaolong *et al.* "Non-local neural networks," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794-7803.
25. Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of IEEE/CVF International Conference on Computer Vision Workshop*, 2019, pp. 1-10.
26. L. Tsung-Yi, G. Priya, G. Ros, H. Kaiming, and D. Piotr, "Focal loss for dense object

detection," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 2980-2988.
27. G. Pavlakos, X. Zhou, and K. Daniilidis, "Ordinal depth supervision for 3D human pose estimation," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7307-7316.
28. T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proceedings of European Conference on Computer Vision*, 2014, pp. 740-755.
29. P. D. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," https://arxiv.org/abs/1412.6980, 2014.
30. A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8026-8037.
31. A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998-6008.
32. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 2961-2969.
33. 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 *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4929-4937.
34. B. J. Frey and D. J. C. MacKay, "A revolution: Belief propagation in graphs with cycles," in *Advances in Neural Information Processing Systems*, 1998, pp. 479-485.

Yi-Yuan Chen (陳奕遠) received his BS degree in Department of Electrical Engineering, National University of Kaohsiung, Taiwan, in 2016, and the MS degree in Department of Computer Science, National Yang Ming Chiao Tung University, Taiwan, in 2021. His research interests include computer vision and deep learning.

Kuochen Wang (王國禎) received the BS degree in Control Engineering from the National Chiao Tung University, Taiwan, in 1978, and the MS and Ph.D. degrees in Electrical Engineering from the University of Arizona in 1986 and 1991, respectively. He is currently a Professor in the Center for Fundamental Science, Kaohsiung Medical University, Taiwan. He was Vice President of the Library and Information Services at this university from August 2019 to August 2021. He was Chair of the Department of Computer Science, National Chiao Tung University from August 2013 to July 2016. He was Director of the Institute of Computer Science and Engineering / Institute of Network Engineering at this university from August 2009 to July 2011. He was Acting / Deputy Director of the Computer and Network Center at this university from June 2007 to July

2009. He was a Visiting Scholar in the Department of Electrical Engineering, University of Washington from July 2001 to February 2002. From 1980 to 1984, he was a Senior Engineer at the Directorate General of Telecommunications in Taiwan. He served in the army as a second lieutenant communication platoon leader from 1978 to 1980. His research interests include artificial intelligence, deep/machine learning, big data analytics, edge/cloud computing, and internet of things.

Hao-Wei Chung (鐘浩瑋) received the M.D. degree from China Medical University, Taichung, Taiwan in 2010. He is currently pursuing the Ph.D. degree at the College of Biological Science and Technology, National Yang Ming Chiao Tung University, Hsinchu, Taiwan. His research interests include action recognition, child development, high risk infants, neonatology and machine learning.

Chien-Chih Chen (陳建智) is a Postdoctoral Research Fellow at Department of Industrial and Information Management, National Cheng Kung University, Taiwan. His current interests focus on machine learning with small data sets. His articles have appeared in Decision Support Systems, Omega, Automation in Construction, Computers and Industrial Engineering, International Journal of Production Research, Neurocomputing, and other publications.

Bohau Huang (黃柏豪) received his MS degree in Computer Science from the National Yang Ming Chiao University in 2021. He received the BS degree in Computer Science from the National Chiao Tung University in 2019. His research interests include machine learning, deep learning, and medical applications.

I-Wei Lu (呂怡緯) received his Ph.D. in Information Management from National Kaohsiung University of Sciences and Technology, Kaohsiung, Taiwan, in 2020. He is currently an Assistant Professor in the Center for Fundamental Science, Kaohsiung Medical University, Kaohsiung, Taiwan. His research interests include information management research, virtual community, online consumer behavior and machine learning. His work in research has been published in the International Journal of Information Management, and Management Review.