H^{3}Net: Irregular Posture Detection by Understanding Human Character and Core Structures

Seungha Noh[1*] seung38231@kyonggi.ac.kr
Kangmin Bae[2] kmbae@etri.re.kr
Yuseok Bae[2] baeys@etri.re.kr
Byong-Dai Lee[1] blee@kyonggi.ac.kr

[1]Kyonggi University, Suwon, Kyonggi, South Korea
[2]ETRI, Daejeon, South Korea

Abstract

This paper proposes H^{3}Net that considers detecting people in irregular postures by utilizing human structures and characters. To handle both features, we introduce two attention modules: 1) Human Structure Attention Module (HSAM), which is introduced to focus on the spatial aspects of a person, and 2) Human Character Attention Module (HCAM), which is designed to address the issue of repetitive appearance. HSAM effectively handles both foreground and background information about a human instance and utilizes keypoints to provide additional guidance to predict irregular postures. Meanwhile, HCAM employs ID information obtained from the tracking head, enriching the posture prediction with high-level semantic information. Furthermore, gathering images of people in irregular postures is a challenging task. Therefore, many conventional datasets consist of images with the same actors simulating varying postures in distinct images. To address this problem, we propose a Human ID Dependent Posture (HID^2) loss that handles repeated instances. The HID^2 loss generates a regularization term by considering duplicated instances to reduce bias. Our experiments demonstrate the effectiveness of H^{3}Net compared to existing algorithms on irregular posture datasets. Furthermore, we show the qualitative results using color-coded masks and bounding boxes. We also provide ablation studies to highlight the significance of our proposed methods.

1. Introduction

Recently, smart surveillance systems [34] and assistance robots [40] have tried to provide detection results of people in irregular posture (*e.g.* lying on the ground). In most cases, people who have fallen on the ground may require immediate medical treatment in a very short time. Due to recent advancements in deep learning, a number of large-scale datasets [3, 7, 30, 32, 62] were proposed to train deep models. These efforts also provoked the creation of large-scale fallen person detection datasets [2, 6]. However, creating a large-scale fallen dataset is a very challenging task. Consequently, most contemporary datasets for fallen detection predominantly employ actors to simulate irregular postures during the image collection process. Due to financial problems, the number of actors is highly limited, which results in a degradation of the generalized posture detectors. In this paper, we refer to this problem as '*limited actor problem*'. While the number of actors is limited, the conventional datasets do not provide ID information to consider the duplication of actors in the training images.

To handle the limited actor problem, we adopt methods from person re-identification and tracking task [66]. The MOT [43] dataset provides bounding box information of all people who appear in the dataset with individual IDs. We adopt this ID information to train the tracker and predict the ID of actors. Furthermore, we utilize this ID information while training the posture predictor to prevent the limited actor problem. By considering ID information, we improved the posture detection performance by limiting the bias that occurred due to the aforementioned problem.

We also adopt structural information while predicting the posture information. A number of studies [19, 25, 33, 51, 64, 71] proved that structural information such as skeleton considerably affects detecting the posture of instances. Therefore, we consider structural as well as ID information while predicting irregular postures. We split the foreground and background of instances by utilizing the segmentation mask of instances. Furthermore, we adopt keypoints prediction with foreground information to detect people in fallen postures. This type of structural information encourages the network to understand the posture of an instance in an irregular posture.

In this paper, we propose H^{3}Net that considers structure and ID information while predicting the posture of

[*] Work done during an internship at ETRI.

(a) Input (b) Segmentation & ID (c) Keypoints (d) Posture

Figure 1. The prediction results of our method on IHP [6] dataset. In Fig. 1a (**Up** and **Down**), the identical instances are located on different backgrounds with different postures. Our approach successfully predicts the segmentation, keypoints, and postures, as well as ID information of instances in diverse scenes. In Fig. 1b, segmentation results are shown in the same color when instances are identified with an identical ID. Also, Fig. 1d demonstrates the posture of each instance through the color of its bounding box (white: normal, red: fallen).

an instance. H^{3}Net handles structural and ID information through the Human Structure Attention Module (HSAM) and the Human Character Attention Module (HCAM). As depicted in Fig. 1, the prediction results of H^{3}Net were shown in order of structure, character, and posture. Furthermore, we also propose a new type of posture loss that considers the limited actor problem. We tested our methods on a large-scale fallen person dataset to show the improvement in posture detection performance.

2. Related Works

2.1. Human Posture Understanding

Many datasets were proposed to understand the posture of a person [2–4, 6, 7, 11, 26, 30, 32, 37, 62] using visual information. One of the methods, such as MPII [3] and COCO [37] datasets, targeted to detect a human pose on various images. To predict more complicated postures, Standford40 [62] and MPHB [7] were proposed with a number of posture classes. For fallen person detection, early methods proposed a number of datasets [1, 5, 13, 41, 67] which are gathered in limited areas. Furthermore, an indoor fallen person detection dataset referred to as IASLAB-RGBD fallen person [4] provided the location and RGBD images of a fallen person in laboratory scenes. Recently, large-scale datasets of fallen person detection were created [2, 6] to understand irregular postures.

Based on these datasets, there were many approaches to understanding the posture of a person. AlphaPose [20, 21, 35], Openpose [8, 9, 52, 59], and Keypoints RCNN [27] were one of the methods that can localize the instance and keypoints simultaneously. While HRNet [18, 53, 58], proposed multi-scale feature space to predict the skeleton of a small person in a bottom-up human pose estimation problem. DEKR [24] utilized adaptive convolution to disentangle the keypoints regeression problems. The development of skeleton prediction provoked a number of studies to predict the action and posture of a person.

STGCN [64] is one of the early methods to predict the action using the keypoints information without any visual information. Based on this research, a number of studies [19, 25, 33, 51, 71] were proposed to predict the action of a person based on pose information. For further understanding of a posture, the interaction between human and object was considered [12]. Detection transformer [10] was used to handle Human-Object Interaction (HOI) problem such as [17, 31, 54, 65, 72]. The interactions that the HOI handles were mostly related to the posture of a person.

2.2. Person Tracking and Re-identification

Multi-Object Tracking (MOT) datasets [23, 43, 55] provides both image-level annotations and inter-frame relationship of instances. MOT datasets provide the ID information throughout the video to inform people's movement between the frames. Furthermore, the advancement of object detection [22, 27, 47, 49, 50, 57] methods provoked the development of tracking algorithms. Based on these datasets and studies, many methods were proposed to handle the tracking problems using detection methods. [42, 46, 56, 66].

Person re-identification also considers the relationship between instances in different frames. A number of works were proposed [16, 28, 39, 61, 68, 70] to study the tem-

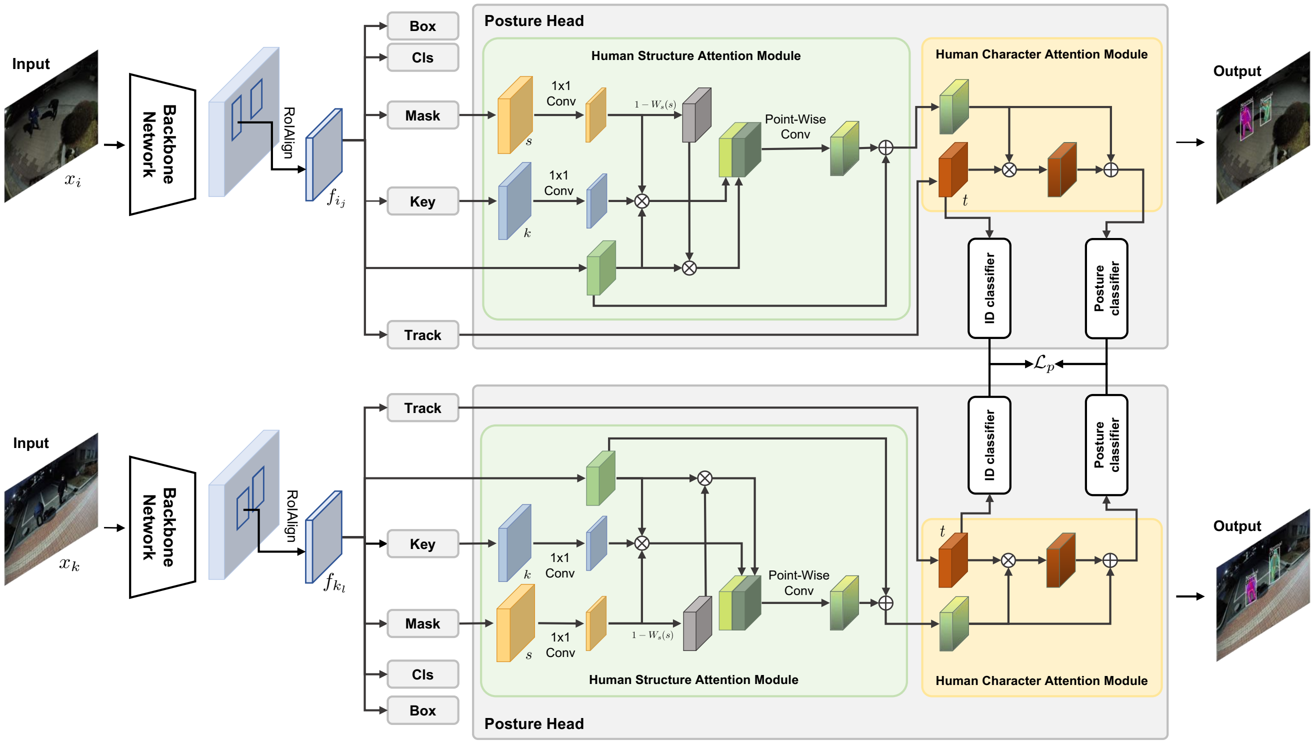


Figure 2. The main architecture of H3Net. Our method consists of 5 heads that predict instances' location, structure, ID, and posture. Our posture heads consist of two attention blocks: 1) HSAM and 2) HCAM. Furthermore, given two different inputs with identical actors, we adopt Human ID Dependent Posture (HID2) loss $\mathcal{L}_p$ while training the posture classifier with predicted ID information.

poral information that can be obtained through a video inputs. As another approach, multi-modal re-identification methods [14, 15, 36, 63] were also proposed to help understand the detailed attributes of a person. Based on these approaches, we propose a network that can track and simultaneously predict an instance's posture. Furthermore, we utilize the ID prediction of a person during the training phase as a regularization term.

3. Methods

We propose H3Net to predict the posture of a person who has fallen. The H3Net has 5 prediction heads: 1) bbox head that estimates the regression problem of the predicted bounding box, 2) mask head that predicts the segmentation mask, 3) keypoints head that predicts the location of human keypoints, 4) tracking head that predicts the appearance embedding of a person, and 5) posture head that considers structural and character information to predict the posture. Additionally, within the posture head, we introduce two attention modules: 1) Human Structural Attention Module (HSAM), and 2) Human Character Attention Module (HCAM). Moreover, we propose the HID2 loss that considers the character information of individuals.

3.1. H3Net

The H3Net considers input image x sampled from dataset $\mathcal{X}$ and posture y sampled from ground truth labels $\mathcal{Y}$. For a given inputs $x_i, x_k \sim \mathcal{X} \times \mathcal{X}$ with ground truth postures y_i and y_k, H3Net obtains n_1 and n_2 regions of interest and extracts features $f_{1_1}, f_{2_1}, ..., f_{i_1}, ..., f_{n_1}$ and $f_{1_2}, f_{2_2}, ..., f_{k_2}, ..., f_{n_2}$ before predicting the postures. As depicted in Fig. 2, the bounding box and classification heads estimate the regression value of location and objectness of f. The mask and keypoints heads predict the structural information of individual instances to consider the pose and region. The tracking head generates appearance embedding of instances to integrate ID information while understanding the posture. Lastly, the posture head predicts the state of a person based on structural and character information inferred from the previous head.

3.2. Human Structure Attention Module

To consider structural information such as posture and appearance of instances, we built the Human Structural Attention Module (HSAM). The HSAM utilizes both mask and keypoints information, which are obtained from the mask head $\mathcal{M}$ and the keypoints head $\mathcal{K}$, respectively, and are defined as follows:

$$\begin{aligned} s &= \mathcal{M}(f) && (1) \\ k &= \mathcal{K}(f), && (2) \end{aligned}$$

where s and k denote the prediction results of the segmentation mask and keypoints heatmap, respectively. Through s, we separate the object region and background region by applying a mask to f. During this process, s initially aligns its channel dimension and spatial size with f by using the convolution layer W_s. The foreground feature l_{fg} and background feature l_{bg} are given as:

$$\begin{aligned} l_{fg} &= f \circ W_s(s) && (3) \\ l_{bg} &= f \circ (1 - W_s(s)), && (4) \end{aligned}$$

where $\circ$ denotes the Hadamard product. To integrate keypoints location, we adopt k to the foreground feature. At this time, k also goes through a convolution layer prior to computation, aiming to align its dimensions and spatial size with l_{fg}. After integrating the keypoints location, we obtain m_{fg} which is given as:

$$m_{fg} = l_{fg} \circ W_k(k). \quad (5)$$

Finally, we concatenate the structured foreground attention feature and background attention feature and compress the concatenated feature using a dimension-based convolution layer. Then, we perform element-wise addition W_{fgbg} with the original feature. As a final output, we perform a concatenation, compression, and addition, which are given as:

$$m_S = W_{fgbg}([m_{fg}, l_{bg}]) + f \quad (6)$$

where m_S denotes the final output of HSAM. Through HSAM, we designed to consider the foreground and background of information before predicting the postures with keypoints locations.

3.3. Human Character Attention Module

The Human Character Attention Module (HCAM) uses outputs from HSAM m_S and identity information obtained from the tracking head $\mathcal{T}$ which is given as:

$$t = \mathcal{T}(f), \quad (7)$$

where t denotes the feature that holds the ID information. Based on t, the ID classifier $\mathcal{I}$ predicts the ID of a human instance C which is given as:

$$C = \mathcal{I}(t). \quad (8)$$

Using t, we obtain ID integrated feature m_{ID}. The m_{ID} is given as:

$$m_{ID} = m_S \circ t. \quad (9)$$

By integrating t, we obtained features that consider not only structural information but also the characters of an instance. As a final output m of HCAM, we perform element-wise addition with the original feature which is given as:

$$m = m_{ID} + f, \quad (10)$$

where m denotes the character information included feature. The m integrates the structural information from m_S and the appearance information of each instance.

3.4. Human ID Dependent Posture (HID^2) Loss

The posture classifier $\mathcal{P}$ takes features obtained from HCAM and classifies the posture p. $\mathcal{P}$ consists of 2 convolution layers and 2 linear layers. p is given as:

$$p = \mathcal{P}(m). \quad (11)$$

Unlike many other datasets, gathering images of humans in irregular postures is not an easy task. Conventional datasets collected various scenes by hiring actors that pretend to have irregular postures. As a result, this method contains a small number of people who are in an irregular state. Therefore, we propose a Human ID Dependent Posture (HID^2) loss that considers duplicated people that appear throughout the whole dataset. The HID^2 loss $\mathcal{L}_p$ is given as:

$$\mathcal{L}_p = \mathbb{E}_i[\mathbb{E}_j[\frac{CE(p_{i_j}, y_{i_j})}{\sum_k \sum_l \mathbb{1}(\mathcal{I}(\mathcal{T}(f_{k_l})) = \mathcal{I}(\mathcal{T}(f_{i_j})))}]], \quad (12)$$

where CE denotes cross entropy loss, and $\mathbb{1}$ denotes indicator function.When f_{i_j} and f_{k_l} turn out to be the feature of identical instances, the total value of the denominator becomes larger. Consequently, when a person appears several times in the training dataset, the weight value of CE decreases to prevent a limited actor problem. As a result, HID^2 loss prevents the bias induced due to multiple occurrences of duplicated actors.

3.5. Training Objective

The final objective of our method is to find optimal parameters that minimize the total loss $\mathcal{L}$. The weighted summation of multiple loss functions is the total loss, which is given as:

$$\begin{aligned} \mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{box} + \mathcal{L}_{segm} \\ + \mathcal{L}_{key} + \lambda_3 \mathcal{L}_{track} + \lambda_4 \mathcal{L}_p, \end{aligned} \quad (13)$$

where λ_1, λ_2, λ_3 and λ_4 are weight parameters. The $\mathcal{L}_{cls}$, $\mathcal{L}_{box}$, $\mathcal{L}_{segm}$, $\mathcal{L}_{track}$, and $\mathcal{L}_{key}$ denote classification, bounding box regression, segmentation, tracking, and keypoints loss independently. Therefore, the final objective is to find the optimal θ^*, which is given as:

$$\theta^* = \arg\min_\theta \mathcal{L}(x; \theta), \quad (14)$$

where θ denotes the network parameters of H^{3}Net.

Dataset	Models		AP	AP50	AP75	APs	APm	APl
IHP [6]	HSENet [6]	bbox	75.768	88.175	83.800	80.099	79.934	73.246
	Iter-E2EDET [69]	bbox	76.555	90.849	84.852	**85.219**	81.534	72.939
	Ours	bbox	**77.542**	**91.030**	**85.883**	51.673	**81.284**	**76.271**
VFP290K [2]	Yolov3 [48]	bbox	59.0	81.3	67.0	-	-	-
	DETR [10]	bbox	60.5	86.8	68.7	-	-	-
	Faster R-CNN [50]	bbox	73.2	87.3	79.9	-	-	-
	Iter-E2EDET [69]	bbox	74.070	89.935	80.909	12.016	66.180	81.248
	Yolov5 [29]	bbox	74.1	83.8	78.4	-	-	-
	DetectoRS [45]	bbox	74.6	86.6	74.6	-	-	-
	Ours	bbox	**74.916**	**88.643**	**81.804**	8.208	69.004	80.435

Table 1. Quantitative results of our methods on IHP [6] and VFP290K [2] datasets. The HSENet [6] and Faster R-CNN [50] utilize ResNet50 with FPN [38] as a backbone network. Our method exhibited improved detection performance bounding box AP metrics for both large-scale fallen person detection datasets.

4. Experiments

4.1. Datasets

To detect people fallen on the ground, we have adopted IHP [6], VFP290K [2], and MOT16 [43] datasets. While VFP290K [2] divided people into two classes: 1) normal and 2) fallen, IHP [6] describes the posture of a person using 6 different classes. Therefore, we combined the non-fallen labels into the 'normal' class, facilitating a comprehensive comparison of detection results with both IHP [6] and VFP290K [2]. The IHP [6] dataset contains $28k$ images with $54k$ instances with keypoints and segmentation labels. While the VFP290K [2] dataset provides $290k$ frames of training images, up to 8 people appear in each frame. The MOT16 [43] dataset provides the ID and location of people in various scenes. We utilized the MOT16 [43] dataset, to train the tracking branch to predict the ID information of the instances.

4.2. Implementation Details

We used ResNet50 with FPN [38] as a backbone network and Detectron2 [60] library in PyTorch [44] was used to implement all networks and classifiers. The bounding box, classification, keypoints, and mask head adopt the network architecture proposed in Mask RCNN [27]. The tracking head adopts Re-ID head of FairMOT [66]. We trained our network on a GPU machine with Intel® Xeon® Gold 6248 @ 2.50GHz CPU and 8 Titan RTX 24GB GPUs. The weight parameters $\lambda_1 = 3$, $\lambda_2 = 2$, $\lambda_3 = 1$, and $\lambda_4 = 0.1$ were used while training the network. We adopted an SGD optimizer with an initial learning rate of 0.02 and weight decay of 0.9.

4.3. Quantitative Results

We have compared our method on both large-scale fallen person detection datasets. As shown in Tab. 1, we measured the detection performance using AP. For detailed information, we also provide the AP measure depending on the threshold of IoUs and the size of the instance. For IHP [6] and VFP290K [2], our model improved up to 2% in AP metric for the overall bounding box detection task. Since the baseline results from VFP290K [2] do not provide AP over the size of instances, we only compare it with varying the threshold of IoU values.

4.4. Qualitative Results

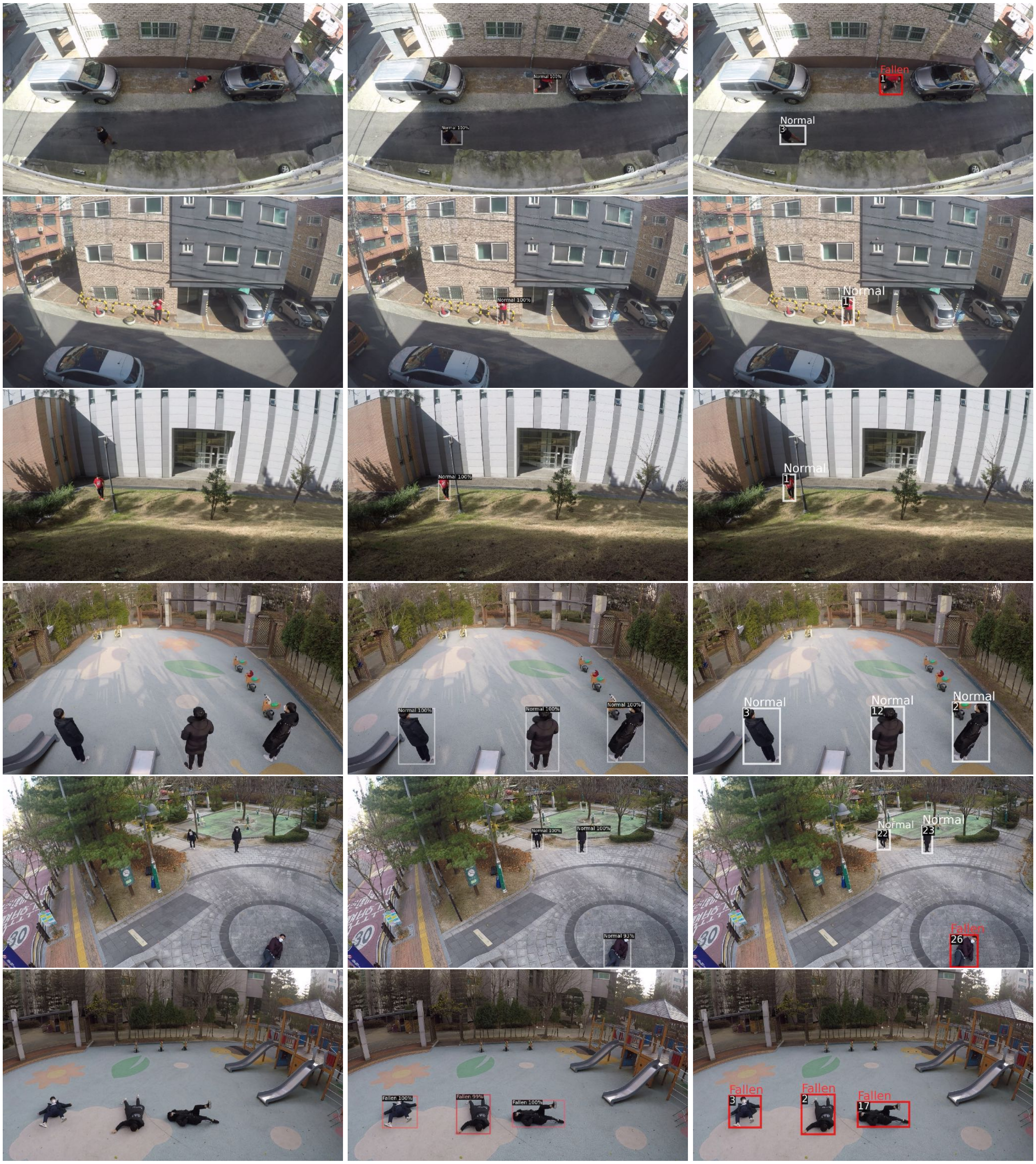
Our method shows improved posture detection results compared to other conventional algorithms. In Figs. 3 and 4, we present our detection results of the fallen person on IHP [6] and VFP290K [2] dataset. Due to the adoption of tracking heads, our method can predict both the posture and ID of people. The ID of each instance is depicted using the color of the predicted segmentation mask. However, in Fig. 3b, the color of the segmentation mask only depicts the predicted posture of instances. The results show that the same tracking ID is assigned to the same person across different scenes, which indicates that our network considers the representation of visual embedding.

4.5. Ablation Studies

We provide the ablation results of our method in Tabs. 2 and 3. We did not include the HID^2 loss, HSAM, and HCAM one by one while training the H^3Net on two training datasets. For experiments on VFP290K [2] dataset, HSAM is not considered due to the nonexistence of segmentation and keypoints annotations. From Tabs. 2 and 3, our method improves the detection performance on both datasets. In Tab. 2, we measured the performance of the networks for 3 different tasks. In particular, we discovered that incorporating two attention modules and HID^2 loss significantly enhanced the AP compared to merely conducting joint training of MOT [43] dataset. In Tab. 3, we present results without HSAM as the VFP290K [2] dataset does not pro-

(a) Inputs (b) HSENet [6] (c) Ours

Figure 3. Qualitative results of IHP [6] dataset. Our method predicts both structural and ID information of each instance. The predicted ID information is depicted through the color of the segmentation mask, and posture is denoted through the color of the bounding box (red: fallen and white: normal).



(a) Inputs (b) Faster R-CNN [50] (c) Ours

Figure 4. Qualitative results of VFP290K [2] dataset. The input images are sampled from the test set of VFP290K [2]. Since VFP290K [2] dataset does not provide structural information, we only predicted the ID of instances. The predicted ID information is written on the top corner of the bounding box, and posture is denoted through the color of the bounding box (red: fallen, and white: normal).

Model	Task	AP	AP50	AP75	APs	APm	APl
Base [6]	bbox	75.768	88.175	83.800	80.099	79.934	73.246
	key	70.378	85.410	77.035	-	77.253	66.755
	segm	70.334	87.724	83.060	53.333	72.177	68.549
+Tracking	bbox	76.607	88.981	84.206	90.000	80.452	73.950
	key	72.579	86.307	78.452	-	78.907	68.863
	segm	71.345	88.886	82.975	18.221	73.186	69.403
+ Tracking + HSAM	bbox	77.187	89.820	85.846	55.149	81.123	73.734
	key	73.739	86.610	80.544	-	80.469	70.334
	segm	70.434	89.452	83.700	22.252	71.752	68.810
+ Tracking + HCAM	bbox	77.288	89.935	**85.943**	81.099	81.727	74.951
	key	73.790	86.715	80.018	-	81.355	70.468
	segm	70.568	89.567	83.577	40.667	71.567	69.340
+ Tracking + HSAM + HCAM	bbox	77.466	89.751	85.695	80.772	81.360	74.823
	key	73.263	86.954	80.136	-	69.862	71.424
	segm	70.577	88.927	83.507	14.155	72.203	68.939
+ Tracking + $\mathcal{L}_p$	bbox	76.851	90.375	85.167	65.050	80.534	74.232
	key	73.393	86.616	78.286	-	78.940	69.927
	segm	72.320	89.510	84.877	40.000	73.951	70.401
Full	bbox	**77.542**	**91.030**	85.883	51.673	81.284	76.271
	key	73.680	87.584	80.087	-	81.220	71.239
	segm	72.658	88.448	83.217	36.465	74.333	71.845

Table 2. The ablation studies of our method on IHP [6] dataset. We measured AP of 3 different tasks: 1) `bbox`, 2) `key`, and 3) `segm`. Based on the segmentation and keypoints predictions, we measured the effects of our methods.

Model		AP	AP50	AP75	APs	APm	APl
Base [2]	bbox	73.2	87.3	79.9	-	-	-
+ Tracking	bbox	74.074	88.517	81.705	9.067	67.462	79.951
+ Tracking + HCAM	bbox	74.377	87.748	80.091	7.867	67.680	79.982
+ Tracking + $\mathcal{L}_p$	bbox	74.473	88.588	80.454	7.981	68.158	80.052
Full	bbox	**74.916**	**88.643**	**81.804**	8.208	69.004	80.435

Table 3. The ablation studies of our method on VFP290K [2] dataset. Since the VFP290K [2] dataset does not provide any structural information, we included character ID-related methods such as HCAM and HID^2 loss while training H^{3}Net. Considering the ID of instances improves the detection performance compared to simple joint training of the tracking head.

vide structural information. Likewise, we observed that our method enhances detection performance compared to conventional multitask training. Our approach effectively mitigates the limited actor problem encountered in training networks for fallen person detection.

5. Conclusions

We proposed an irregular posture detection network referred to as H^{3}Net, which integrates HID^2 loss along with two specialized attention modules: 1) Human Structure Attention Module (HSAM) and 2) Human Character Attention Module (HCAM). The HSAM deals with structural information such as segmentation mask and keypoints location, while the HCAM focuses on ID prediction, encompassing character information for each instance. To address the prevalent limited actor problem in large-scale fallen person datasets, we have also incorporated the use of HID^2 loss, which utilizes the predicted ID. This approach ensures that H^{3}Net effectively overcomes the existing limitations in detecting irregular human postures in these datasets.

Acknowledgments

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2014-3-00123, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis, and No.2022-0-00124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities).

References

[1] Kripesh Adhikari, Hamid Bouchachia, and Hammadi Nait-Charif. Activity Recognition for Indoor Fall Detection using Convolutional Neural Network. In *IAPR International Conference on Machine Vision Applications (MVA)*, 2017. 2

[2] Jaeju An, Jeongho Kim, Hanbeen Lee, Jinbeom Kim, Junhyung Kang, Minha Kim, Saebyeol Shin, Minha Kim, Donghee Hong, and Simon S. Woo. VFP290K: A Large-Scale Benchmark Dataset for Vision-based Fallen Person Detection. In *NeurIPS Datasets and Benchmarks Track (Round 2)*, 2021. 1, 2, 5, 7, 8

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *CVPR*, 2014. 1, 2

[4] Morris Antonello, Marco Carraro, Marco Pierobon, and Emanuele Menegatti. Fast and Robust detection of fallen people from a mobile robot. In *IROS*, 2017. 2

[5] Edouard Auvinet, Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. Multiple cameras fall data set. *Technical report 1350*, 2011. 2

[6] Kangmin Bae, Kimin Yun, Jungchan Cho, and Yuseok Bae. The Dataset and Baseline Models to Detect Human Postural States Robustly against Irregular Postures. In *AVSS*, 2021. 1, 2, 5, 6, 8

[7] Yawei Cai and Xiaoyang Tan. Weakly supervised human body detection under arbitrary poses. In *ICIP*, 2016. 1, 2

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*, 2017. 2

[9] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE TPAMI*, 2019. 2

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *ECCV*, 2020. 2, 5

[11] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. HICO: A Benchmark for Recognizing Human-Object Interactions in Images. In *ICCV*, 2015. 2

[12] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to Detect Human-Object Interactions. In *WACV*, 2018. 2

[13] Imen Charfi, Johel Miteran, Julien Dubois, Mohamed Atri, and Rached Tourki. Definition and Performance Evaluation of a Robust SVM Based Fall Detection Solution. In *International Conference on Signal Image Technology and Internet Based Systems*, 2012. 2

[14] Cuiqun Chen, Mang Ye, Meibin Qi, and Bo Du. Sketch Transformer: Asymmetrical Disentanglement Learning from Dynamic Synthesis. In *ACM MM*, 2022. 3

[15] Cuiqun Chen, Mang Ye, and Ding Jiang. Towards Modality-Agnostic Person Re-identification with Descriptive Query. In *CVPR*, 2023. 3

[16] Guangyi Chen, Yongming Rao, Jiwen Lu, and Jie Zhou. Temporal Coherence or Temporal Motion: Which Is More Critical for Video-Based Person Re-identification? In *ECCV*, 2020. 2

[17] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating HOI Detection As Adaptive Set Prediction. In *CVPR*, 2021. 2

[18] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. In *CVPR*, 2020. 2

[19] Hyung-Gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. InfoGCN: Representation Learning for Human Skeleton-based Action Recognition. In *CVPR*, 2022. 1, 2

[20] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional Multi-person Pose Estimation. In *ICCV*, 2017. 2

[21] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE TPAMI*, 2022. 2

[22] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding YOLO Series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2

[23] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *CVPR*, 2012. 2

[24] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-Up Human Pose Estimation Via Disentangled Keypoint Regression. In *CVPR*, 2021. 2

[25] Pallabi Ghosh, Yi Yao, Larry S. Davis, and Ajay Divakaran. Stacked Spatio-Temporal Graph Convolutional Networks for Action Segmentation. In *WACV*, 2020. 1, 2

[26] Saurabh Gupta and Jitendra Malik. Visual Semantic Role Labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2

[27] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 2, 5

[28] Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, and Shiguang Shan. BiCnet-TKS: Learning Efficient Spatial-Temporal Representation for Video Person Re-Identification. In *CVPR*, 2021. 2

[29] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, ChristopherSTAN, Liu Changyu, Laughing, tkianai, Adam Hogan, lorenzomammana, yxNONG, AlexWang1900, Laurentiu Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, Francisco Ingham, Frederik, Guilhen, Hatovix, Jake Poznanski, Jiacong Fang, Lijun Yu, changyu98, Mingyu Wang, Naman Gupta, Osama Akhtar, PetrDvoracek, and Prashant Rai. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements, 2020. 5

[30] S. Johnson and M. Everingham. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *BMVC*, 2010. 1, 2

[31] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. HOTR: End-to-End Human-Object Interaction Detection With Transformers. In *CVPR*, 2021. 2

[32] Kyung-Rae Kim, Whan Choi, Yeong Jun Koh, Seong-Gyun Jeong, and Chang-Su Kim. Instance-Level Future Motion Estimation in a Single Image Based on Ordinal Regression. In *ICCV*, 2019. 1, 2

[33] Sunoh Kim, Kimin Yun, Jongyoul Park, and Jin Young Choi. Skeleton-Based Action Recognition of People Handling Objects. In *WACV*, 2019. 1, 2

[34] Sergio Lafuente-Arroyo, Pilar Martín-Martín, Cristian Iglesias-Iglesias, Saturnino Maldonado-Bascón, and Francisco Javier Acevedo-Rodríguez. RGB camera-based fallen person detection system embedded on a mobile platform. *Expert Systems with Applications*, 197:116715, 2022. 1

[35] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient Crowded Scenes Pose Estimation and a New Benchmark. In *CVPR*, 2019. 2

[36] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person Search With Natural Language Description. In *CVPR*, 2017. 3

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 2

[38] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 5

[39] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality Aware Network for Set to Set Recognition. In *CVPR*, 2017. 2

[40] Saturnino Maldonado-Bascón, Cristian Iglesias-Iglesias, Pilar Martín-Martín, and Sergio Lafuente-Arroyo. Fallen people detection capabilities using assistive robot. *Electronics*, 8(9):915, 2019. 1

[41] Georgios Mastorakis and Dimitrios Makris. Fall detection system using Kinect's infrared sensor. *Journal of Real-time Image Processing*, 2014. 2

[42] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *CVPR*, 2022. 2

[43] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A Benchmark for Multi-Object Tracking. *arXiv:1603.00831 [cs]*, 2016. arXiv: 1603.00831. 1, 2, 5

[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*. 2019. 5

[45] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. DetectoRS: Detecting Objects With Recursive Feature Pyramid and Switchable Atrous Convolution. In *CVPR*, 2021. 5

[46] Zheng Qin, Sanping Zhou, Le Wang, Jinghai Duan, Gang Hua, and Wei Tang. MotionTrack: Learning Robust Short-Term and Long-Term Motions for Multi-Object Tracking. In *CVPR*, 2023. 2

[47] Joseph Redmon and Ali Farhadi. YOLO9000: Better, Faster, Stronger. In *CVPR*, 2017. 2

[48] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. *arXiv*, 2018. 5

[49] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*, 2016. 2

[50] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*, 2015. 2, 5, 7

[51] Fumiaki Sato, Ryo Hachiuma, and Taiki Sekii. Prompt-Guided Zero-Shot Anomaly Action Recognition Using Pretrained Deep Skeleton Features. In *CVPR*, 2023. 1, 2

[52] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *CVPR*, 2017. 2

[53] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep High-Resolution Representation Learning for Human Pose Estimation. In *CVPR*, 2019. 2

[54] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. QPIC: Query-Based Pairwise Human-Object Interaction Detection with Image-Wide Contextual Information. In *CVPR*, 2021. 2

[55] Paul Voigtlaender, Michael Krause, Aljoša Ošep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTS: Multi-Object Tracking and Segmentation. In *CVPR*, 2019. 2

[56] Paul Voigtlaender, Jonathan Luiten, Philip H.S. Torr, and Bastian Leibe. Siam R-CNN: Visual Tracking by Re-Detection. In *CVPR*, 2020. 2

[57] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-YOLOv4: Scaling Cross Stage Partial Network. In *CVPR*, 2021. 2

[58] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE TPAMI*, 2019. 2

[59] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional Pose Machines. In *CVPR*, 2016. 2

[60] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 5

[61] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly Attentive Spatial-Temporal Pooling Networks for Video-based Person Re-Identification. In *ICCV*, 2017. 2

[62] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human Action Recognition by Learning bases of Action Attributes and Parts. In *ICCV*, 2011. 1, 2

[63] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel Augmented Joint Learning for Visible-Infrared Recognition. In *ICCV*, 2021. 3

[64] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In *AAAI*, 2018. 1, 2

[65] Frederic Z. Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring Predicate Visual Context for Detecting Human-Object Interactions. In *ICCV*, 2023. 2

[66] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the Fairness of Detection and Re-identification in Multiple Object Tracking. *International Journal of Computer Vision*, 129, 2021. 1, 2, 5

[67] Zhong Zhang, Christopher Conly, and Vassilis Athitsos. Evaluating Depth-Based Computer Vision Methods for Fall Detection under Occlusions. In *Advances in Visual Computing*, 2014. 2

[68] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xiansheng Hua. Attribute-Driven Feature Disentangling and Temporal Aggregation for Video Person Re-Identification. In *CVPR*, 2019. 2

[69] Anlin Zheng, Yuang Zhang, Xiangyu Zhang, Xiaojuan Qi, and Jian Sun. Progressive End-to-End Object Detection in Crowded Scenes. In *CVPR*, 2022. 5

[70] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. MARS: A Video Benchmark for Large-Scale Person Re-Identification. In *ECCV*, 2016. 2

[71] Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Learning Discriminative Representations for Skeleton Based Action Recognition. In *CVPR*, 2023. 1, 2

[72] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, and Jian Sun. End-to-End Human Object Interaction Detection With HOI Transformer. In *CVPR*, 2021. 2