Towards Accurate Human Pose Estimation in Videos of Crowded Scenes

Shuning Chang[*]
Li Yuan[*]
yuanli@u.nus.edu
National University of Singapore

Xuecheng Nie
YiTu Technology

Ziyuan Huang
National University of Singapore

Yichen Zhou
YiTu Technology

Yunpeng Chen
YiTu Technology

Jiashi Feng
National University of Singapore

Shuicheng Yan
YiTu Technology

ABSTRACT

Video-based human pose estimation in crowed scenes is a challenging problem due to occlusion, motion blur, scale variation and viewpoint change, etc. Prior approaches always fail to deal with this problem because of (1) lacking of usage of temporal information; (2) lacking of training data in crowded scenes. In this paper, we focus on improving human pose estimation in videos of crowded scenes from the perspectives of exploiting temporal context and collecting new data. In particular, we first follow the top-down strategy to detect persons and perform single-person pose estimation for each frame. Then, we refine the frame-based pose estimation with temporal contexts deriving from the optical-flow. Specifically, for one frame, we forward the historical poses from the previous frames and backward the future poses from the subsequent frames to current frame, leading to stable and accurate human pose estimation in videos. In addition, we mine new data of similar scenes to HIE dataset from the Internet for improving the diversity of training set. In this way, our model achieves best performance on 7 out of 13 videos and 56.33 average w_AP on test dataset of HIE challenge.

KEYWORDS

pose estimation, object detection, human in events

ACM Reference Format:
Shuning Chang, Li Yuan, Xuecheng Nie, Ziyuan Huang, Yichen Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. 2020. Towards Accurate Human Pose Estimation in Videos of Crowded Scenes . In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3394171.3416299

1 INTRODUCTION

Human pose estimation is important for many computer vision applications, including human action recognition, human-computer interaction, and video surveillance. Due to the viewpoint variance, appearance variance and cluttered background, pose estimation is a very challenging task for large scale image and video datasets. Recently, significant progress has been made in this area [20]. However, the pose estimation in complex events [14] is still relatively new and a challenging problem. In this challenge of pose estimation on crowed scenes and complex events, we propose to obtain the pose of single images based pose estimation method, which can be applied to each video frame to get an initial pose estimation, and a further refinement through frames can be applied to make the pose estimation consistent and more accurate.

The general pipeline for the pose estimation method that we used can be divided into two parts, human detection and pose estimation, respectively. First, we use a human detection method in crowd scenes to detect the bounding box of highly-overlapping human instances in the detection phase. In the second step, we perform pose estimation on every box by two state-of-the-art single-person pose estimation models [20, 24]. During the pose estimation phase, we propose a optical flow smoothing algorithm to refine our pose predictions. The framework of our approach is shown in Figure 1.

Since the problem is treated as a two-stage problem to be tackled one by one, each module will be introduced separately. The following of the report is organized as follows: Sec. 2 investigates the common detection methods on HIE2020 challenge and also introduces our detailed method and experiments on human detection; Sec. 3 introduces the pose estimation as well as the final pose generation process. Sec. 4 introduces the experiments and training details of pose estimation. Finally, Sec. 5 concludes the report.

2 HUMAN DETECTION

The first step of human pose tracking is to detect the bounding boxes of person. As no validation set in HIE dataset [14], we split the original training set as a new training set and validation set. Two splitting strategies are tried: splitting by image frames (5k for validation, 27k for training) and splitting by videos. We found that splitting by image will cause over-fitting and far away from the data distribution of the testing set. So we adopt the video-splitting strategy and split video 3,7,8 and 17 as the validation set and the

[*] Authors contributed equally to this work. Work done during internship at YiTu Technology.

MM '20, October 12–16, 2020, Seattle, WA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3416299

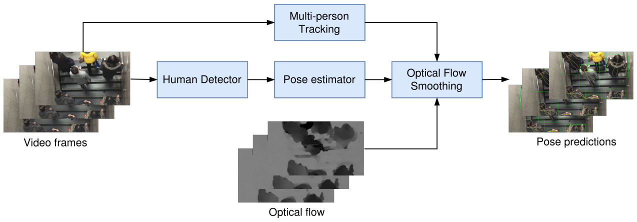

Figure 1: The framework of our approach.

rest videos as train data, in which 5.7k image frames for validation and the reset 27k images frames for training. Based on the train and validation set, we can conduct detection experiments on HIE. All the performance of models is tested by two metrics, Averaged Precision (AP) and MMR [3]. AP reflects both the precision and recall ratios of the detection results; MMR is the log-average Miss Rate on False Positive Per Image (FPPI) in [0.01, 100], is commonly used in pedestrian detection. MR is very sensitive to false positives (FPs), especially FPs with high confidences will significantly harm the MMR ratio. Larger AP and smaller MMR indicates better performance.

2.1 Common Detection Frameworks

There are mainly two different types of common detection frameworks: one-stage (unified) frameworks [15–17] and two-stage (region-based) framework [5–7, 18]. Since RCNN [6] has been proposed, the two-stage detection methods have been widely adopted or modified [1, 10, 12, 18, 22, 23, 27, 30, 32]. Normally, the one-stage frameworks can run in real-time but with the cost of a drop in accuracy compared with two-stage frameworks, so we mainly adopt two-stage frameworks on HIE dataset.

We first investigate the performance of different detection backbone and framework on HIE dataset, including backbone: ResNet152 [8], ResNeXt101 [26] and SeNet154 [9], and different framework: Faster-RCNN [18], Cascade R-CNN [1], and Feature-Pyramid Networks (FPN) [11]. The experimental results on different backbone and methods are given in Table 1. The baseline model is Faster RCNN with ResNet50, and we search hyper-parameters on the baseline model then apply to the larger backbone. From table 1, we can find that the better backbone (ResNet152 and ResNeXt101) and combining advanced methods (Cascade and FPN) can improve the detection performance, but the SENet154 does not get better performance than ResNet152 even it has superior classification performance on ImageNet. So in our final detection solution, we only adopt ResNet152 and ResNeXt101 as the backbone.

Table 1: Performance comparison (AP and mMR) among different detection backbone and methods on HIE dataset.

Methods or Modules	AP (%)	MMR (%)
Baseline (ResNet50 + Faster RCNN)	61.68	74.01
ResNet152 + Faster RCNN	67.32	68.17
ResNet152 + Faster RCNN + FPN	69.77	64.83
SENet154 + Faster RCNN + FPN	65.77	68.46
ResNeXt101 + Faster RCNN + FPN	69.53	63.91
ResNeXt101 + Cascade RCNN + FPN	71.32	61.58
ResNet152 + Cascade RCNN + FPN	71.06	62.55

2.2 Extra Data for Human Detection

In the original train data, there are 764k person bounding boxes in 19 videos with 32.9k frames, and the testing set contains 13 videos with 15.1k frames. Considering the limited number of videos and duplicated image frames, the diversity of train data is not enough. And the train data and test data have many different scenes, thus extra data is crucial for training a superior detection model. Here we investigate the effects of different human detection dataset on HIE, including all the person images in COCO (COCO person, 64k images with 262k boxes) [13], CityPerson (2.9k image with 19k boxes) [31], CrowndHuman (15k images with 339k boxes) [19] and self-collected data (2k images with 30k boxes). We investigate the effects on different data based on Faster-RCNN with ResNet50 as the backbone. The experimental results are shown in Table 2. We can find that the CrowdHuman dataset achieves the largest improvement compared with other datasets, because the CrowdHuman is the most similar scenes with HIE, and both of the two datasets contain plenty of crowded scenes. COCO person contains two times of images than HIE train data, but merging the COCO person does not bring significant improvement and suffer more than three times train time, thus we only merge HIE with CrowdHuman and self-collected data to take a trade-off between detection performance and train time.

2.3 Detection in Crowded Scenes

As there are lots of crowded scenes in HIE2020 dataset, the highly-overlapped instances are hard to detect for the current detection framework. We apply a method aiming to predict instances in crowded scenes [2], named as “CrowdDet”. The key idea of CrowdDet is to let each proposal predict a set of correlated instances rather than a single one as the previous detection method. The CrowdDet includes three main contributions for crowded-scenes detection: (1) an EMD loss to minimize the set distance between the two sets of proposals [21]; (2). A refine module that takes the combination of predictions and the proposal feature as input, then performs a second round of predicting. (3). Set NMS, it will skip normal NMS suppression when two bounding boxes come from the same proposal, which has been proved works in crowded detection; We conduct experiments to test the three parts on HIE2020 dataset, and the results are shown in Table 3. Based on the results in the Table, we can find that the three parts do improve the performance in crowded detection. Meanwhile, we apply KD regularization [28, 29] in the class's logits of the detection model, which can consistently improve the detection results by 0.5%-1.4%.

Finally, based on the above analysis, we train two detection models on HIE by combining extra data with the crowded detection framework: (1). ResNet152 + Cascade RCNN + extra data + emd loss + refine module + set NMS + KD regularization, whose AP is 83.21; (2). ResNeXt101 + Cascade RCNN + extra data + emd loss + refine module + set NMS + KD regularization, whose AP is 83.78; Then two models are fused with weights 1:1.

3 POSE ESTIMATION

In this section, we will introduce the networks we used to generate pose estimation and the optical flow smoothing algorithm serving for smoothing the pose predictions.

Table 2: The effects of using extra data for human detection on HIE dataset.

Validation set	AP (%)	MMR (%)
HIE data	61.68	74.01
HIE + COCO person	65.83	69.75
HIE + CityPerson	63.71	67.43
HIE + CrowdHuman	**78.22**	**58.33**
HIE + self-collected data	**69.39**	**60.82**
HIE + CrowdHuman + COCO + CityPerson	78.53	58.63
HIE + CrowdHuman + self-collected data	**81.03**	**55.58**
HIE + all extra data	81.36	55.17

Table 3: Detection in Crowded Scenes on HIE dataset.

Validation set	AP (%)	MMR (%)
ResNet50 + Faster RCNN + extra data	81.36	55.17
+ emd loss	81.73	53.20
+ refine module	81.96	50.85
+ set NMS	**82.05**	**49.63**

3.1 Single-person Pose Estimators

We adopt two state-of-the-art single-person pose estimation models, HRNet [20] and SimpleNet [25], as our basic networks to generate pose predictions. Different from general high-to-low and low-to-high pattern, HRNet can maintain the high-resolution representations through the whole process and fuse multi-resolution representations simultaneously. SimpleNet is a simple and effective model, which just consists of a backbone network, ResNet in our work, declining the resolution of the feature map, and several deconvolutional layers producing the pose predictions. Additionally, for SimpleNet, we plug an FPN [10] structure in it to strengthen the performance of small person instances. Finally, we fuse the results of two models by averaging their heatmaps.

3.2 Optical Flow Smoothing

This task is based on videos, so the temporal information is a potentially available condition. Moreover, most people's actions in this dataset are not bouncing, just simple standing, sitting, and walking, so the poses from the same person are similar between adjacent frames. However, our single models cannot capture the temporal relationship. To solve this issue, we design an optical flow smoothing algorithm to smooth our pose predictions.

We propose to smooth the current frame from the previous frame and the next frame by optical flow which is often expressed for temporal information. Given one human instance with joints coordinates set $\widehat{J_i^k}$ in current frame I^k, first we compute J_i^{k-1} in frame I^{k-1} and the optical flow field $F_{k-1 \longrightarrow k}$ between frame I^{k-1} and I^k, then we can estimate the current frame joints coordinates set $\hat{J}_i^{k-1 \longrightarrow k}$ in frame I^k by propagating the joints coordinates set J_i^{k-1} according to $F_{k-1 \longrightarrow k}$. Specifically, for each joint location (x, y) in J_i^{k-1}, the propagated joint location will be $(x + \delta x, y + \delta y$, where δx, δy are the flow field values at joint location (x, y). Similarly, we can estimate the current frame $\hat{J}_i^{k+1 \longrightarrow k}$ from the next frame in the same way. Finally, we obtain the final predicted J_i^k as follows:

Table 4: The top-3 results of HIE2020 testing set. The evaluation metric is w-AP(%).To compare with the results for each video, we highlight the best results by red color and highlight the second one by blue color. Our approach achieves the best results on the vast majority of videos

Video Name	First Place	Ours	Third Place
hm_in_waiting_hall	64.5796	65.8270	58.5896
hm_in_bus	56.0834	55.4518	50.6453
hm_in_dining_room2	22.2609	25.9449	23.5640
hm_in_lab2	72.7162	70.3300	69.9452
hm_in_subway_station	41.5776	47.2997	41.3249
hm_in_passage	88.9244	90.1478	86.5233
hm_in_fighting4	57.1941	59.8902	56.3970
hm_in_shopping_mall3	61.2707	62.7075	60.9024
hm_in_restaurant	58.2427	67.4902	64.3151
hm_in_accident	53.1889	56.9365	54.6401
hm_in_stair3	47.7152	49.1054	49.6768
hm_in_crossroad	75.5781	75.7640	73.8597
hm_in_robbery	51.2827	52.2467	51.4776
Weighted Average	57.5091	56.3375	55.1719

$$J_i^k = \alpha \cdot \hat{J}_i^{k-1 \longrightarrow k} + \alpha \cdot \hat{J}_i^{k+1 \longrightarrow k} + (1 - 2\alpha) \cdot \hat{J}_i^k, \qquad (1)$$

where the α is used to weighted sum the three terms.

The bottleneck of our method is how to track the same person in the adjacent frames. Some traditional work applies bounding box IoU (Intersection-over-Union) or pose similarity to link instances [25]. However, there are numerous ultra crowded scenes in this dataset, which leads to severe occlusion and overlap, so the traditional methods would be problematic. Different previous work, we use a person Re-ID (person Re-identification) model to extract features to compute similarity. Compared with other methods, the Re-ID features focus on human appearance more, therefore, they are more suitable for this dataset. To verify our inference, we submit our tracking result to track 1 (private) server and achieve 61.0951% on MOTA metric, which demonstrates the effect of our Re-ID features.

For the whole procedure of our optical flow smoothing algorithm, first, we utilize our person tracking model using Re-ID features to generate the person ID. Then, if the same IDs exist in the previous and next frames and their confidence scores are higher than a threshold, we will use Eqn 1 to smooth our pose estimation.

4 EXPERIMENTS

4.1 Extra Data for Human Pose

The original official training set contains about 660.5K annotated poses. Considering that large-scale similar data exist due to frame-wise annotation, it is necessary to collect extra data to improve the performance. The extra training data we used come from two aspects: (1) We fuse three mainstreaming public pose estimation datasets, COCO, MPII, and AI Challenger, into our training data. The COCO dataset contains over 250k person instances labeled with 17 key points. The MPII dataset consists of 25K images including over 40K person instances with annotated 16 body joints. The AI

Table 5: Performance evaluation of different components in our method on the HIE testing set.

HRNet	SimpleNet	Multi-scale Evaluation	Multi-scale Input	Extra Data	Optical Flow Smoothing	w-AP(%)
✓						52.45
✓		✓				52.90
✓				✓		53.82
✓	✓	✓		✓		55.52
✓	✓	✓	✓	✓		56.04
✓	✓	✓	✓	✓	✓	56.34

Challenger dataset is composed of about 700K person instances with annotated 14 body joints. Since the annotated key points in these datasets are not totally overlapped with official labels, for each dataset we use respective overlapped key points for training. (2) Self-collected data with similar scenes are merged into our training set. The number of poses is not over 30K, which is far less than the official training data.

All the extra data are randomly merged into the official training set. We do not explore more complicate data fusion strategy.

4.2 Training Details

We extend the human detection box in height or width to a fix 4 : 3 aspect ratio, and then crop the box from the image, which is resized to a fixed size, 256 × 192 or 384 × 288. The data augmentation includes random rotation($[-45°, 45°]$), random scale$(0.65, 1.35)$, and flipping. Half body data augmentation is also applied. The network of HRNet used is HRNet-W48 and the backbone of SimpleNet used is ResNet152. We implement our training using PyTorch.

4.3 Testing Details

We adopt a multi-scale evaluation during testing. Specifically, we rescale the detection box to obtain new bounding boxes with different scales, then crop them to the original size and flip them to acquire their flipped counterparts. The generated boxes are feed into the network to produce heatmaps. We average those heatmaps and search the highest response to obtain the locations of key points. The scale factors used are 0.7, 1.0, and 1.3. Moreover, it is easy to suffer redundancy and wrong boxes in the complex and crowded scenes. We apply Pose NMS [4] to eliminate similar and low-confidence redundancies.

4.4 Results

The top-3 results of HIE2020 testing set are shown in Table 4. From our results for each video, we can see that our method achieves significant performance in the regular and high-resolution videos, such as "hm_in_passage" and "hm_in_crossroad". Our method performs poorly in the video with crowded scenes and low quality, *e.g.*, we only get 25.94% on the ultra crowded video "hm_in_dining_room2", which is much lower than other videos. Our results for each video have remarkable performance. Even if compared with the first place, except "hm_in_bus" and "hm_in_lab2" are totally lower than them by 3%, we achieve better performance in the rest videos. However, our weighted average result is 1.2% lower than the winner. We analyse the possible reason is to exceed false positive predictions in our results. The false positive predictions are from two aspects: first, redundancy bounding boxes cause redundancy pose predictions; second, some small person instances are not involved in evaluation but we produce their poses.

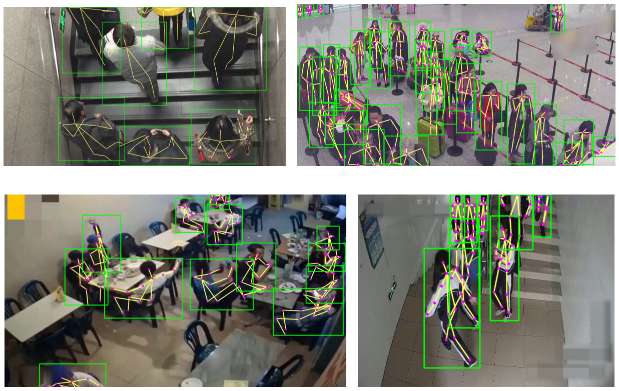

Figure 2: Example pose estimation results on the HIE2020 test set.

We visualize some example of our pose estimation results in Figure 2, which illustrates our approach can produce accurate pose predictions in the complex and crowded scenes.

4.5 Ablation Study

In order to verify the performance of our components, we have done extensive experiments. The experiment results are shown in Table 5. Note that "Multi-scale Input" means training multiple groups of parameters by changing input size and fusing their results during testing. For each ablation experiment, if there is a ✓in the "Multi-scale Input" cell, the results is obtained by fusing input size 256×192 and 384 × 288; otherwise, the input size is just 256 × 192. Extra data significantly boost our results by about 1.4%, implying the effectiveness of large-scale data. Two fusion methods, model fusion between HRNet and SimpleNet and multi-scale input also improves our result tremendously by 1.7% and 1.5% respectively. Our post-processing algorithm, optical flow smoothing, can enhance the results by 0.3%, which shows that it is effective.

5 CONCLUSION

In this paper, we illustrate the approach we used in the HIE2020 Challenger pose estimation track. We adopt a top-down approach to address this complex and crowded scene issue. First, for human detection problems in crowed scenes, we add extra data to overcome the overfitting problem and apply one proposal for multiple predictions to relieve the difficulty of detecting highly-overlapping instances. Then, we apply our effective single-person pose estimation model to generate accurate pose predictions. To utilize temporal information, we design an optical flow smoothing algorithm to post-process our results.

REFERENCES

[1] Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6154–6162.

[2] Xuangeng Chu, Anlin Zheng, Xiangyu Zhang, and Jian Sun. 2020. Detection in Crowded Scenes: One Proposal, Multiple Predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12214–12223.

[3] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. 2011. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence* 34, 4 (2011), 743–761.

[4] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. RMPE: Regional Multi-Person Pose Estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[5] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[9] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[10] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

[11] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

[12] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[14] Weiyao Lin, Huabin Liu, Shizhan Liu, Yuxi Li, Guo-Jun Qi, Rui Qian, Tao Wang, Nicu Sebe, Ning Xu, Hongkai Xiong, et al. 2020. Human in Events: A Large-Scale Benchmark for Human-centric Video Analysis in Complex Events. *arXiv preprint arXiv:2005.04490* (2020).

[15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

[16] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[17] Joseph Redmon and Ali Farhadi. 2016. YOLO9000: Better, Faster, Stronger. *arXiv preprint arXiv:1612.08242* (2016).

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[19] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. 2018. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123* (2018).

[20] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5693–5703.

[21] Siyu Tang, Mykhaylo Andriluka, and Bernt Schiele. 2014. Detection and tracking of occluded people. *International Journal of Computer Vision* 110, 1 (2014), 58–69.

[22] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. 2019. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4933–4942.

[23] Tao Wang, Xiaopeng Zhang, Li Yuan, and Jiashi Feng. 2019. Few-shot adaptive faster r-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7173–7182.

[24] Bin Xiao, Haiping Wu, and Yichen Wei. 2018. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*. 466–481.

[25] Bin Xiao, Haiping Wu, and Yichen Wei. 2018. Simple Baselines for Human Pose Estimation and Tracking. In *The European Conference on Computer Vision (ECCV)*.

[26] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1492–1500.

[27] Li Yuan, Eng Hock Francis Tay, Ping Li, and Jiashi Feng. 2019. Unsupervised Video Summarization with Cycle-consistent Adversarial LSTM Networks. *IEEE Transactions on Multimedia* (2019).

[28] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. 2019. Revisit knowledge distillation: a teacher-free framework. *arXiv preprint arXiv:1909.11723* (2019).

[29] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. 2020. Revisiting Knowledge Distillation via Label Smoothing Regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3903–3911.

[30] Li Yuan, Francis EH Tay, Ping Li, Li Zhou, and Jiashi Feng. 2019. Cycle-sum: Cycle-consistent adversarial lstm networks for unsupervised video summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9143–9150.

[31] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. 2017. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3213–3221.

[32] Li Zhou, Jian Zhao, Jianshu Li, Li Yuan, and Jiashi Feng. 2018. Object Relation Detection Based on One-shot Learning. *arXiv preprint arXiv:1807.05857* (2018).