Nearby-person Occlusion Data Augmentation for Human Pose Estimation with Non-extra Annotations

Yucheng Chen, Mingyi He*, Yuchao Dai
Northwestern Polytechnical University, Xian 710129, China
* Email address: myhe@nwpu.edu.cn (Mingyi He)

***Abstract*—Human pose estimation has made significant progress with deep learning techniques, while the estimation of occlusion keypoints is still an unsolved problem. One important reason comes from the insufficiency of the existing benchmark datasets, such as imbalanced body keypoints annotation and lack of occluded training samples. To address this problem, we propose Nearby-person Occlusion Data Augmentation (NODA), a method that provides synthetic nearby-person occlusion images by only utilizing existing annotations. First, we generate rough mask of human bodies with keypoints annotation to build a foreground human body pool. Then, one foreground human body crop is randomly sampled and properly placed over the training human body to synthesis nearby-person occlusion training images. The proposed data augmentation method is easy to implement and deploy to any other methods. Extensive experimental results on MPII benchmark demonstrate the effectiveness of our method with Simple and HRNet as the backbone models. Especially on easily-confusable joints, our method makes significant improvement.**

***Index Terms*—Human Pose Estimation; Nearby-person Occlusion; Data Augmentation; Deep learning**

I. Introduction

Human pose estimation (HPE) is a research spot of computer vision tasks aiming to estimate pixel-wise keypoints of people on the images [1], which plays an important role in a variety of high-level vision tasks, such as action recognition [2]–[4], action detection [5], human tracking [6], etc. Due to the booming of deep learning in recent years [7]–[10], the performance of human pose estimation has made significant achievements [11]–[15]. However, there are still some unsolved challenges, such as occlusion.

So far, the existing benchmark dataset such as MPII [16] is labeled with the visibility of keypoints, but few studies have used these annotations. Meanwhile, the insufficiency of imbalanced body keypoints annotation and lack of occluded training samples also influence the estimation performance. As show in Fig. 1, the performance of each keypoint is highly correlated with the amount of training data, and the estimation accuracy of occluded keypoints are much lower than visible keypoints.

Existing methods to handle occlusion with data augmentation can be divided into two types. One way is to synthesis occluded images by pasting over parts from other object datasets [17] or off-the-shelf segmentation models [18]. Another way is information dropping which is to erase body region such as random erasing [19], Cutout [20], hide-and-seek (HaS) [21] and GridMask [22]. These methods either require additional annotation data, or the generated image is quite different from the real image.

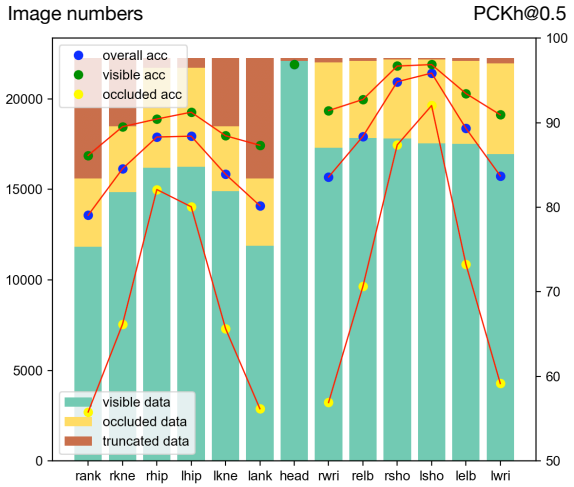

Fig. 1: MPII analysis based on Simple [23] with ResNet50 as backbone. The keypoints annotations are imbalanced, and the occluded training samples are insufficient. The performance of occluded keypoints is inferior. Here "ank", "kne", "wri", "elb" and "sho" respectively indicate "ankle", "knee", "wrist", "elbow" and "shoulder". The "r" and "l" stand for "right" and "left".

In this article, we propose a novel nearby-person occlusion data augmentation approach to synthesize training images for human pose estimation. First, we generate rough mask of human body with keypoints annotation to build a foreground human body pool. Then, one foreground human body crop is randomly sampled and properly placed over the training human body to synthesis nearby-person occlusion training images. In summary, our main contributions are two-fold:

- We propose a novel nearby-person occlusion data augmentation (NODA) approach to synthesize more occluded training images with non-extra annotations. The synthesized occluded images properly balance the training set.
- We comprehensively evaluate our method on benchmark dataset MPII and demonstrate the effectiveness of our method with different levels of mask.

II. Related Work

A. Human Pose Estimation

Recently, pose estimation using DCNNs has shown superior performance. DeepPose [11] firstly attempted to apply an

(a) Keypoint-Masking [24] (b) Parsing-based Data Augmentation [18] (c) Information Dropping Methods [25]

Fig. 2: **Top: Illustration of nearby-person occlusion data augmentation.** The rough masks are obtained with existing annotations. With the occlusion overlay strategy, the occluded images are generated for training. **Bottom: Other related methods.** (a) and (b) belong to synthetic occlusion way. (c) belongs to information dropping way.

AlexNet-like deep neural network to learn keypoints coordinates from full images in a very straightforward manner. [12] proposed a heatmap representation for each keypoint and largely improved the spatial generalization. Following the heatmap-based framework, various methods [13]–[15], [23] focused on designing the structure of the network and indeed achieved significant improvement. However there are still some unsolved challenges, such as occlusion. In this work, taking advantage of the well-designed network structure Simple [23] and HRNet [15], we propose a novel data augmentation solution to further improve the performance of human pose estimation.

B. Occlusion Data Augmentation

Pasting over parts and information dropping of images are two types of data augmentation widely used in image classification [26], object detection [27], person re-identification [28]. In human pose estimation, Keypoint-Masking [24] augment images by copying background patches over some of the keypoints. [17] synthesis occluded images by pasting over parts from other object datasets. Bin et al. [18] proposed adversarial semantic data augmentation with off-the-shelf segmentation models. [25] summarized information dropping methods [19]–[22] and proposed to increase training time to improve performance. Different from the existing data augmentation strategies, we propose a novel nearby-person data augmentation scheme which takes advantage of the non-extra human keypoint annotations to obtain the mask of whole body rather than only body parts, other objects or noisy image patches.

(a) (b) (c) (d)

Fig. 3: **Three levels of mask** (a) Orignal Image, (b) Box-level Mask, (c) Cardboard-level Mask, (d) GrabCut-level Mask.

III. Methodology

As show in Fig. 2 is our proposed nearby-person occlusion data augmentation method. Firstly, we utilize the keypoint-level annotation to generate rough mask of human bodies of each training sample to build a foreground human body pool. Then, for each training image, one foreground human body crop is randomly sampled and properly placed over the training human body to synthesis nearby-person occlusion training image.

A. Rough Mask Generation

In order to verify the effectiveness of the proposed data augmentation method, we generate foreground human body crops on three different levels of mask, providing more accurate human masks in turn, shown in Fig. 3. The ablation studies Sec. IV-C show that more precise human segmentation will lead to better results.

1) Box-level Mask Generation: As show in (b) of Fig. 3, with human body keypoint labels, we can obtain a compact bounding boxes. Then a rough bounding box is cropped with

TABLE I: **Ablation studies.** The performance of the visible and the occluded keypoints are listed in the parentheses (visible result, occluded ressult). Here "Box.", "Card.", "Crab.", "Smo." and "Res." respectively indicate "Box-level mask", "Cardboard-level mask", "GrabCut-level mask", "Edge smoothing" and "Occlusion rescaling".

Method	Box.	Card.	Crab.	Smo.	Res.	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	PCKh@0.5
Simple [23]						96.4 (96.8, 0)	95.3 (96.7, 89.2)	89.0 (93.5, 71.2)	83.2 (91.0, 57.5)	88.4 (90.7, 81.2)	84.0 (89.6, 64.2)	79.6 (87.0, 57.2)	88.5 (92.8, 70.8)
Simple	√					96.8 (+0.4, 0)	95.6 (+0.1, +1.2)	89.7 (+0.9, -0.1)	84.1 (+0.6, +1.9)	89.5 (+1.0, +1.2)	85.3 (+1.4, +0.9)	80.6 (+1.0, +1.2)	89.3(+0.8) (+0.7, +1.1)
Simple		√				96.7 (+0.4, 0)	95.8 (+0.3, +1.3)	89.3 (+0.4, -0.2)	84.4 (+0.9, +2.2)	89.0 (+0.6, +0.6)	84.8 (+1.0, +0.5)	80.9 (+0.6, +3.4)	89.2(+0.7) (+0.6, +1.3)
Simple			√			96.8 (+0.5, 0)	95.5 (-0.1, +0.9)	89.6 (+0.8, -0.3)	84.3 (+0.4, +3.2)	88.8 (+0.6, -0.2)	84.3 (+0.7, -0.9)	79.6 (-0.4, +1.5)	89.0(+0.5) (+0.4, +0.7)
Simple			√	√		97.1 (+0.7, 0)	95.8 (+0.2, +1.4)	89.5 (+0.8, -0.6)	84.7 (+0.9, +3.8)	89.3 (+1.0, +0.5)	84.8 (+1.0, +0.6)	80.0 (+0.2, +0.9)	89.3(+0.8) (+0.7, +1.1)
Simple			√	√	√	96.8 (+0.4, 0)	95.9 (+0.2, +1.9)	90.0 (+0.8, +1.8)	84.4 (+0.8, +2.6)	89.2 (+0.7, +1.1)	85.5 (+1.6, +1.1)	80.3 (+0.2, +2.7)	**89.4(+0.9) (+0.7, +1.8)**
Simple	√	√	√	√	√	96.8 (+0.4, 0)	95.6 (+0.2, +0.7)	89.4 (+0.7, -0.8)	84.6 (+0.7, +4.0)	89.2 (+0.9, +0.5)	84.9 (+1.1, +0.2)	80.5 (+0.5, +2.6)	89.3(+0.8) (+0.6, +1.2)

a proper scale expansion. The box-level mask crops contain many background information of the bodies.

2) Cardboard-level Mask Generation: By utilizing the cardboard model [29], we can crop each body part in the type of bounding box, thus to get the whole body mask. As show in (c) of Fig. 3, the cardboard-level mask crops contain less background information, while may loss a small amount of body information due to imprecise segmentation.

3) GrabCut-level Mask Generation: GrabCut [30] is an efficient interactive foreground/background segmentation method based on graph cuts. There are two steps to apply GrabCut to generate body masks: (1) finding the smallest bounding box of the human body region from the keypoint annotation and (2) generating a body mask based on the bounding box, clear foreground region (the keypoints and corresponding connection) and clear background region (the information outside of the bounding box). The result of GrabCut is show in (d) of Fig. 3

B. Occlusion Overlay Strategy

To synthesis reasonable nearby-person occlusion training image, the degree of occlusion needs to be appropriate. The coverage can neither be occluded too severely nor too slightly. Following the keypoint annotation distribution of MPII, we set the following coverage rules: (1) Several keypoints are not allowed to be covered, such as head, neck, etc. (2) The four limbs (arms and legs) should be occluded at least one keypoint.

The occlusion overlay strategy is shown in Fig. 2. The red point and yellow point indicate the center of target people body and occlusion people body, respectively. For each training image, the region between two dotted circles centered on the center of the yellow point is the candidate region to place the foreground nearby people. We set the radius of target people as R which is half of the diagonal of the bounding box. Then, the red point can be confirmed based on a random radius ($1/3R$ - R) and a random rotation angle ($0°$ - $360°$).

IV. Experiments

A. Dataset and Evaluation Protocol

We evaluate our method on a representative benchmark Max Planck Institute for Informatics (MPII) human pose dataset [16]. The MPII dataset includes around 25k images with poses of 40k people annotated with 2D locations of 16 keypoints. Following [23], 2975 samples are taken as a validation set. Our models are trained on a subset of MPII training set and evaluate on the validation set. We have evaluated our method on MPII dataset with the percentage of correct keypoints (PCKh) which measures the localization accuracy of the predicted keypoints. After measuring distance between the groundtruth keypoints and predicted keypoints, PCKh counts the number of keypoints that are within selected distance thresholds normalized by head size. PCKh@0.5 indicates the threshold of distance is 0.5 times head size.

B. Implementation Details

In order to show the effectiveness of our proposed data augmentation method, we have set Simple [23] and HRNet [15] as our baseline models. Both methods are top-down methods, and the basic structure of both methods is widely used in human pose estimation.

In case of Simple [23] , we adopt ResNet-50 and ResNet-152 models and input image resolution of (256,256) for MPII. We use data augmentations such as rescaling(30%), rotation(40 degrees) and flip. Following [23] the Adam optimizer [31] is used. For training details, the base learning rate is 1e-3. It drops to 1e-4 at 90 epochs and 1e-5 at 120 epochs. There are 140 epochs in total. Mini-batch size is 64. The Simple is initialized with weight of pre-trained model on public-released ImageNet [32].

In case of HRNet, we adopt the HRNet-w32 model and input image resolution of (256,256) for MPII. Similarly, data augmentations include random rescaling([0.65, 1.35]), random rotation($\pm 45°$) and flip. We employ the Adam optimizer [31]. The learning schedule follows the setting [15]. The base

TABLE II: **Comparisons on MPII val set.** The visible and the occluded keypoints performance are listed in the parentheses (visible result, occluded result).

Method	Backbone	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	PCKh@0.5
Simple [23]	ResNet-50	96.4 (96.8, 0)	95.3 (96.7, 89.2)	89.0 (93.5, 71.2)	83.2 (91.0, 57.5)	88.4 (90.7, 81.2)	84.0 (89.6, 64.2)	79.6 (87.0, 57.2)	88.5 (92.8, 70.8)
Simple+NODA	ResNet-50	96.8 (+0.4, 0)	95.9 (+0.2, +1.9)	90.0 (+0.8, +1.8)	84.4 (+0.8, +2.6)	89.2 (+0.7, +1.1)	85.5 (+1.6, +1.1)	80.3 (+0.2, +2.7)	**89.4(+0.9) (+0.7, +1.8)**
Simple [23]	ResNet-152	97.0 (97.0, 0.0)	95.9 (97.2, 90.8)	90.0 (94.3, 72.9)	85.0 (92.1, 60.7)	89.2 (91.7, 81.7)	85.3 (89.9, 67.6)	81.3 (87.8, 59.4)	89.6 (93.5, 72.8)
Simple+NODA	ResNet-152	97.0 (+0.0, 0)	95.8 (-0.2, -0.3)	90.2 (+0.1, +0.3)	85.1 (+0.2, +0.2)	89.8 (+0.4, +1.5)	86.1 (+0.9, +0.5)	82.4 (+1.3, +0.7)	**90.0(+0.4) (+0.3, +0.5)**
HRNet [15]	HRNet-w32	97.1 (97.1, 0.0)	95.8 (97.1, 90.4)	90.3 (94.6, 73.0)	85.0 (92.9, 58.2)	89.0 (91.5, 81.8)	87.0 (91.2, 70.7)	83.1 (89.6, 61.5)	90.0 (93.9, 73.1)
HRNet+NODA	HRNet-w32	97.4 (+0.3, 0.0)	96.3 (+0.4, +1.4)	91.6 (+0.9, +3.4)	86.9 (+0.8, +5.7)	89.6 (+0.5, +0.7)	87.4 (+0.1, +1.6)	83.8 (-0.2, +3.4)	**90.9(+0.9) (+0.4, +2.7)**

learning rate is set as 1e-3, and is dropped to 1e-4 and 1e-5 at the 170th and 200th epochs, respectively. The training process is terminated within 210 epochs. Mini-batch size is 64. The HRNet is initialized with weight of pre-trained model on public-released ImageNet [32].

C. Ablation Studies

In order to verify the effectiveness of different hyper-parameters in the method, we employ Simple [23] with ResNet-50 as backbone to do ablation studies. Half of training samples are randomly occluded with our data augmentation method. Both the PCKh of the visible and the occluded keypoints are are shown.

1) Mask type comparison: First of all, we evaluate the performance of three types of mask. Note only the occlusion overlay strategy is used. The occlusion people mask is rescaled as 1.0 to the target people. As shown in Table I, the performance of the visible and the occluded keypoints are listed in the parentheses. All three types of occlusion data augmentation are helpful to improve the performance, especially the occlusion accuracy. The box-level mask contains more background information may lead to more realistic occlusion. The cardboard-level and GrabCut-level mask show more influence on the keypoints at the end of limbs such as wrist and ankle.

2) Edge smoothing comparison: In order to generate more realistic occlusion, we reduce the opacity of the GrabCut-level mask along the border for smoother blending [17]. The results of GrabCut-level mask in Table I are all improved with edge smoothing.

3) Occlusion rescale comparison: We randomly rescale the occlusion human body size in the range of (0.8-1.2) to provide more diversified occlusion. The ablation experiment is done on smoothed GrabCut-level mask shown in Table I with improvement.

4) NODA: Combine all three types of mask, edge smoothing, occlusion rescaling, the model reports 89.3% less than the model trained only with CrabCut-level mask. The additional background information and unreal human edges may not help to the results. Therefore, the NODA includes GrabCut-level mask overlay, edge smoothing, occlusion rescaling.

D. Results

The performance of proposed method on MPII val set are listed in Table II. With ResNet-50 as backbone model, the improvement is around 0.9% PCKh@0.5 by using NODA. Deeper baseline with ResNet-152 still brings 0.4% improvement with NODA. Based on HRNet-w32, our model with NODA achieves 0.5% increase. The main improvement comes from the increasing of the occlusion point estimation accuracy. In particular, the occlusion keypoints subset show considerable improvement especially on wrist, knee and ankle which are considered as the most challenging keypoints. The consistency in performance improvement proves both the universal effectiveness of the proposed nearby-person occlusion data augmentation method.

E. Qualitative Results

Fig. 4 visualizes some estimated results to qualitatively showcase the efficiency of the proposed method. From left to right are ground truths, estimated results by HRNet-w32 [15] and estimated results with NODA. The yellow circles marks the wrong prediction. We can observe that the predictions of HRNet-w32 are confused by occlusion. By applying NODA to augment training images, we improves the performance of the original HRNet-w32 in the occluded challenging cases, providing more reasonable results.

V. Conclusions

In this paper, we propose a novel nearby-person occlusion data augmentation (NODA) approach to synthesize training images for human pose estimation. First, we generate rough mask of human body with keypoints annotation to build a foreground human body pool. Then, one foreground human body crop is randomly sampled and properly placed over the training human body to synthesis nearby-person occlusion training images. Without using any extra annotations, we comprehensively evaluate our method on benchmark dataset MPII with Simple and HRNet as backbone models demonstrating the effectiveness of our method.

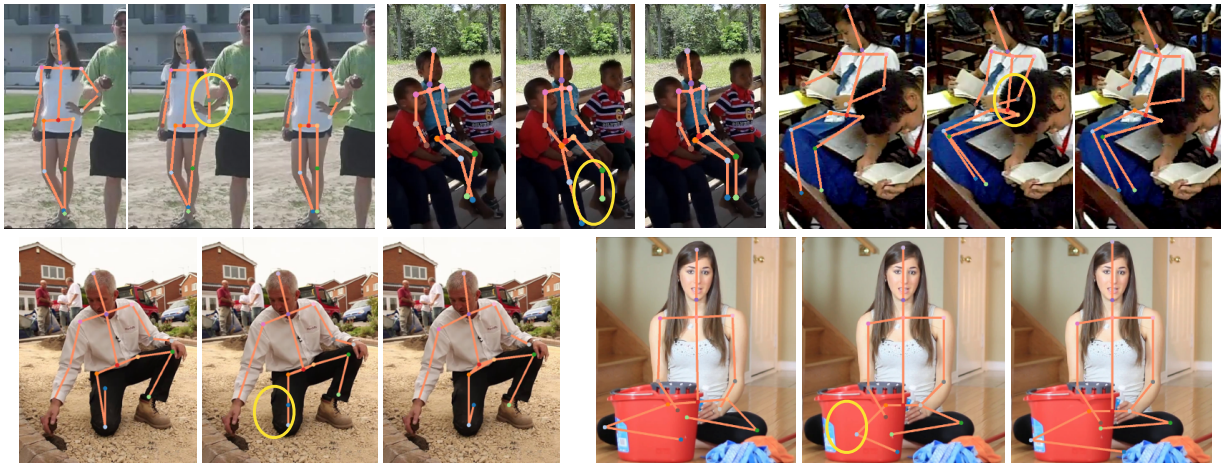

Fig. 4: Visualization of results. From left to right are ground truths, original estimated results and estimated results with NODA.

ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China (61671387, 61871325 and 62001396).

REFERENCES

[1] Yucheng Chen, Yingli Tian, and Mingyi He, "Monocular human pose estimation: A survey of deep learning-based methods," *Computer Vision and Image Understanding*, vol. 192, pp. 102897, 2020. 1

[2] Bo Li, Yuchao Dai, Xuelian Cheng, Huahui Chen, Yi Lin, and Mingyi He, "Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn," in *International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2017, pp. 601–604. 1

[3] Diogo C Luvizon, David Picard, and Hedi Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018, pp. 5137–5146. 1

[4] Bo Li, Mingyi He, Yuchao Dai, Xuelian Cheng, and Yucheng Chen, "3d skeleton based action recognition by video-domain translation-scale invariant mapping and multi-scale dilated cnn," *Multimedia Tools and Applications*, vol. 77, no. 17, pp. 22901–22921, 2018. 1

[5] Bo Li, Huahui Chen, Yucheng Chen, Yuchao Dai, and Mingyi He, "Skeleton boxes: Solving skeleton based action detection with a single deep convolutional neural network," in *International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2017, pp. 613–616. 1

[6] Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, and Bernt Schiele, "Arttrack: Articulated multi-person tracking in the wild," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 6457–6465. 1

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NIPS)*, vol. 25, pp. 1097–1105, 2012. 1

[8] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 1

[9] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9. 1

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. 1

[11] Alexander Toshev and Christian Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1653–1660. 1

[12] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," *Advances in Neural Information Processing Systems (NIPS)*, vol. 27, pp. 1799–1807, 2014. 1, 2

[13] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh, "Convolutional pose machines," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4724–4732. 1, 2

[14] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7291–7299. 1, 2

[15] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, "Deep high-resolution representation learning for human pose estimation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5693–5703. 1, 2, 3, 4

[16] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3686–3693. 1, 3

[17] István Sárándi, Timm Linder, Kai O Arras, and Bastian Leibe, "Synthetic occlusion augmentation with volumetric heatmaps for the 2018 eccv posetrack challenge on 3d human pose estimation," *arXiv preprint arXiv:1809.04987*, 2018. 1, 2, 4

[18] Yanrui Bin, Xuan Cao, Xinya Chen, Yanhao Ge, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Changxin Gao, and Nong Sang, "Adversarial semantic data augmentation for human pose estimation," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 606–622. 1, 2

[19] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang, "Random erasing data augmentation," in *AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 13001–13008. 1, 2

[20] Terrance DeVries and Graham W Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017. 1, 2

[21] Krishna Kumar Singh, Hao Yu, Aron Sarmasi, Gautam Pradeep, and Yong Jae Lee, "Hide-and-seek: A data augmentation technique for weakly-supervised localization and beyond," *arXiv preprint arXiv:1811.02545*, 2018. 1, 2

[22] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia, "Gridmask data augmentation," *arXiv preprint arXiv:2001.04086*, 2020. 1, 2

[23] Bin Xiao, Haiping Wu, and Yichen Wei, "Simple baselines for human pose estimation and tracking," in *European Conference on Computer Vision (ECCV)*. Springer, 2018, pp. 466–481. 1, 2, 3, 4

[24] Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu, "Multi-scale structure-aware network for human pose estimation," in *European Conference on Computer Vision (ECCV)*. Springer, 2018, pp. 713–728. 2

[25] Junjie Huang, Zheng Zhu, Guan Huang, and Dalong Du, "Aid: Pushing the performance boundary of human pose estimation with information dropping augmentation," *arXiv preprint arXiv:2008.07139*, 2020. 2

[26] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018, pp. 7132–7141. 2

[27] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Z Li, and Xudong Zou, "Pedhunter: Occlusion robust pedestrian detector in crowded scenes," in *AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 10639–10646. 2

[28] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Conference on Computer Vision and Pattern Recognition Workshops(CVPRW)*. IEEE, 2019. 2

[29] Shanon X Ju, Michael J Black, and Yaser Yacoob, "Cardboard people: A parameterized model of articulated image motion," in *the Second International Conference on Automatic Face and Gesture Recognition*. IEEE, 1996, pp. 38–44. 3

[30] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake, "Grabcut interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004. 3

[31] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 3

[32] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255. 3, 4