Tilting at windmills: Data augmentation for deep pose estimation does not help with occlusions

Rafal Pytel
Computer Vision Lab
Delft University of Technology

Osman Semih Kayhan
Computer Vision Lab
Delft University of Technology

Jan C. van Gemert
Computer Vision Lab
Delft University of Technology

***Abstract*—Occlusion degrades the performance of human pose estimation. In this paper, we introduce targeted keypoint and body part occlusion attacks. The effects of the attacks are systematically analyzed on the best performing methods. In addition, we propose occlusion specific data augmentation techniques against keypoint and part attacks. Our extensive experiments show that human pose estimation methods are not robust to occlusion and data augmentation does not solve the occlusion problems.** [1]

I. Introduction

Human Pose Estimation is the task of localizing anatomical keypoints such as eyes, hips, knees and localizing body-parts like head, limbs, corpus, etc., with many applications in segmentation [24], [25], [52], action recognition [28], [30], [40], pose tracking [14], [54], gait recognition [39], [44], autonomous driving [12], [32], [50], elderly monitoring [10], [31] and social behaviour analysis [22], [48]. All these applications rely on correct and robust pose estimation. In this paper we investigate the robustness of human pose estimation methods to a natural and common effect: Occlusions.

Occlusions are common and occur frequently in the wild as for example by a random object, another person [15], and self-occlusion [18]. Prior works address occlusion in a general way and exploits segmentation [32] or depth information [33]. Where [36] evaluates robustness with image and domain-agnostic universal perturbations. In contrast, we systematically analyze targeted occlusion attacks not only for keypoints, but also for and body parts and investigate the sensitivity of pose estimation to occlusion attacks.

A promising solution to occlusions is data augmentation, which is practically a default setting for deep learning applications [37] where image flipping, rotation, and scaling offer endless data variations [6], [37], [45]. As such, regional dropout and mixup methods improve the generalization performance of image classification [9], [16], [41], [46], [49], [55], [59], [60], object localization and detection [7], [11], [38] and segmentation [13]. In pose estimation, [19] applies region based augmentation by exchanging a single keypoint patch with a random background patch. More recent approaches [42], [53] use half-body augmentation wherewith the presence of more than 8 keypoints, by choosing upper or lower body keypoints. We implement systematic data augmentation methods for occlusion for keypoint and body parts to investigate how data augmentation can remedy occlusion attacks.

[1]For the code and the extended version:
https://github.com/rpytel1/occlusion-vs-data-augmentations

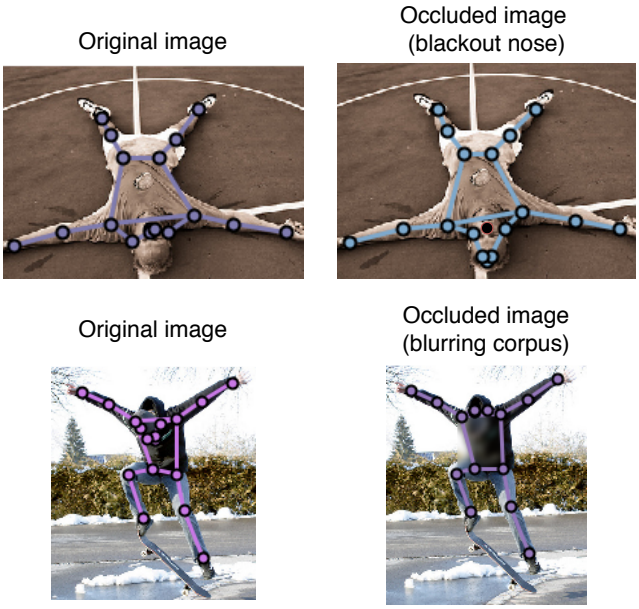
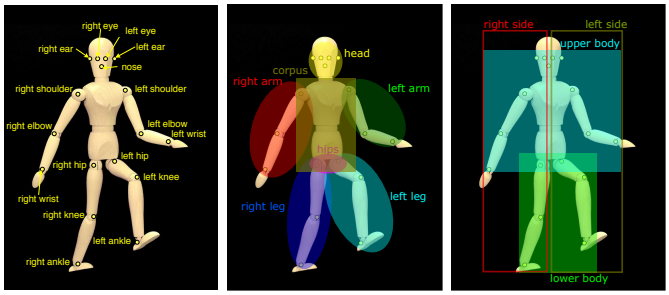


Fig. 1: Qualitative example how HRNet-32 [42] predictions change after keypoint blackout on the nose (first row) and part blurring on the corpus (second row). For both examples keypoints change for head, nose, eyes and ears.

We have the following contributions: First, we conduct a structured investigation on the occlusion problem of pose estimation and introduce occlusion attacks. Second, we investigate occlusion-based data augmentation methods. Third, we show that data augmentation does not provide robustness to occlusion attacks.

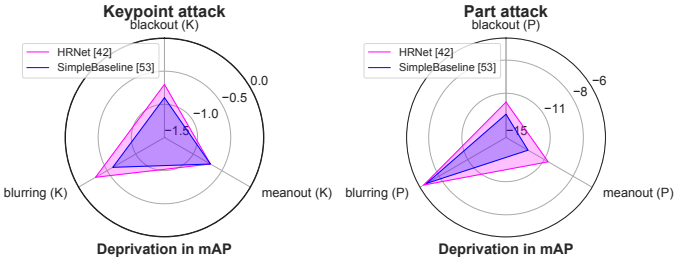
II. Related work

Human Pose Estimation. Deep learning methods in human pose estimation can be divided into 2 categories: bottom-up and top-down. Bottom-up approaches [3], [6], [21], firstly localize identity-free keypoints and then group them into person instances. Top-down approaches [5], [29], [42], [53] firstly detect a person in the image and then perform a single person estimation within the bounding box. The top-down approaches achieve the state of the art results on various multi-person benchmarks such as COCO [26], MPII [1]. Within top-down



(a) COCO keypoints annotations. (b) Part mapping for smaller parts. (c) Part mapping for larger parts.

Fig. 2: Visualization of keypoint annotations in COCO dataset and proposed part mapping.



(a) (b)

Fig. 3: Robustness comparison of HRNet [42] and SimpleBaseline [53] against (a) keypoint and (b) part occlusion attacks. HRNet is more robust against both attacks, yet both attacks drop performance, where part attacks deteriorate more.

approaches 2 categories can be distinguished: regressing direct location of each keypoint [4], [47] and keypoint heatmaps estimation [8], [29], [42], [51], [53] followed by choosing the locations with the highest heat values as the keypoints. The best performing methods on COCO keypoint challenge use a cascade network [5], [23] to improve keypoint prediction. The 'SimpleBaseline' [53] proposes simple but effective improvement by adding few deconvolutional layers to enlarge the resolution of output features. HRNet [42] which is built from multiple branches can produce high-resolution feature maps with rich semantics and performs well on COCO. Some works advance performance of HRNet via improvement over standard encoding and decoding of heatmaps [58] and basing data processing on the unit length instead of pixels [17] with an additional off-set strategy for encoding and decoding. Because of their good accuracy and wide adaptation, we focus on top-down methods, HRNet and SimpleBaseline and bottom-up approach Higher HRNet.

Occlusion in pose estimation. Occlusion in pose estimation is an under-studied problem. In [36] analyses of occlusions are done for deep pose estimators by domain-agnostic universal perturbations. More recently, attempts to solve the occlusion problem in pose estimation are suggested via the usage of segmentation of occluded parts [32] and depth of in an image [33]. OcclusionNet [34] predicts occluded keypoints via graph-neural networks yet it is applied only on vehicles. Different from these methods, in our paper we introduce keypoint occlusion attacks and body part occlusion attacks and give a structured analysis of occlusion on human pose estimation.

Data augmentation. Data augmentation is a strong, simple and popular approach to increase model robustness. Removing part of the image improves generalization of image classification [9], [55], [60] and object localization-detection [7], [11], [38]. Mixup [16], [46], [59] approaches which create a combination of two images are often used in image classification. [13][57] combine regional dropout and MixUp methods for image segmentation [13] and image classification [57] task. [19] proposes a cutmix-like approach where a small patch from the background is pasted on the single keypoint or vice versa. For the human pose estimation methods [4], [51], [56], scaling, rotation and flipping is commonly used as data augmentation. Random cropping is also used in bottom-up approaches [3], [6], [21]. More recent top-down approaches [5], [42], [53] employ the usage of half body transform by a probability of 0.3 choosing either upper or lower body keypoints. We introduce and evaluate new data augmentation methods for keypoint and for body parts specifically designed against occlusion attacks for human pose estimation.

III. Sensitivity to Occlusion Attacks

We investigate the effect of occlusion attacks on MS COCO dataset [26]. COCO contains challenging images with the unconstrained environment, different body scales, variety of human poses and occlusion patterns. The dataset contains over 200k images with 250k person instances labelled with 17 keypoints. Models are trained on COCO train2017 datasets which includes 57k images and 150k person instances. The evaluation is done on val2017 set which contains 5k images.

The occlusion attack experiments are conducted with HRNet [42] and Simple Baseline [53] for two aspects: (i) keypoint attacks, where the occlusion area is a centred circle on the chosen keypoint, (ii) body part attacks, where the occlusion area is the minimum rectangle covering all keypoints of a chosen part. The COCO keypoints and the proposed groups of body parts can be seen in Figure 2. For the analyses, COCO pretrained HRNet and Simple Baseline are evaluated by the performance of the network against keypoint and part occlusion attacks on COCO validation set.

HRNet and SimpleBaseline produce heatmap instead of predicting direct single location for each keypoint. The ground truth heatmaps are generated by using 2D Gaussian of size 13x13. Thus, as a default, we choose the size of the occlusion circle with a radius of 6 pixels for keypoint attacks to cover the keypoint heatmap. We have 3 different keypoint attacks: (i) Gaussian blur (blurring) attack, (ii) attack by filling with black pixels (blackout), (iii) attack by filling with a mean intensity value of a given image (meanout).

Body parts occlusion attacks are designed to draw a minimum rectangle which covers all the keypoints of a chosen part. Similar to the keypoint attacks, we have 3 different part

attacks which are applied to the occlusion area: blurring with the kernel size 31 and sigma 5, blackout and meanout. These attacks can be applied on both small parts such as head, arms, hips and larger parts like upper body, lower body, left and right side (Figure 2 b and c).

We compare HRNet and Simple Baseline according to their robustness to keypoint and part occlusion attacks. Figure 3 shows that both attacks are quite successful as occlusion causes the performance to drop. HRNet is more robust against keypoint and part occlusion attacks. For further analyses, we only use HRNet as a baseline for our investigations.

A. How sensitive to key point occlusion attacks?

First, we analyze the effect of the occlusion size on the average performance of the pose estimator on all keypoints. Figure 4 indicates that pose estimator performance is inversely proportional to the occlusion size and blurring, blackout, and meanout attacks on average perform similarly. The size of the occlusion decreases the average performance of the estimator by approximately 3% when the radius of the occlusion circle is chosen as 18 pixels.

Second, we show the class-specific performance drops for each individual keypoints for each attack. In Figure 5, attacking nose causes serious loss in mAP, almost 5% for blackout, 4.4% for meanout and 1.2% for blurring. The empirical results indicate that **the nose** is the most important keypoint since the occlusion of the nose causes notable performance drop. After the nose, each eye influences the performance of other keypoints mostly by approximately 1% with each occlusion attack. Keypoints from less densely annotated places like ankles or wrists are the least influential.

If we check the analysis of the reduced accuracy per keypoint for the case of attacking nose (Figure 6a), the most affected keypoints are the ones within close distance, which are eyes and ears due to being a part of the head. Interestingly, occluding nose affects the performance of the left eye estimation more than occluding the left eye itself, respectively by approximately 10% and 5% (Figure 6a, 6b). If we investigate per keypoint performance for occluding left ankle, it can be seen that the deprivation is by several magnitudes smaller than in case of the nose or left eye occlusions. From the observation of the analyses, it can be drawn that HRNet [42] is not robust to keypoint occlusion attacks.

B. How sensitive to part occlusion attacks?

We analyze the effect of the part occlusion attacks on each body parts given in Figure 2. Attacking the upper body, left and right sides influence the overall performance the most, by more than 44%, 24% and 24% with blackout attack respectively since these three parts include the majority of the keypoints (Figure 7). When we examine keypoint-specific accuracy drops for the remaining keypoints of the upper body, it is clear that blackout is the most influential attack, with a drop of almost 3% for left and right ankle (Figure 8a). If we investigate per-keypoint behaviour for the corpus (Figure 8b), we observe significant degradation of the performance on all the keypoints,

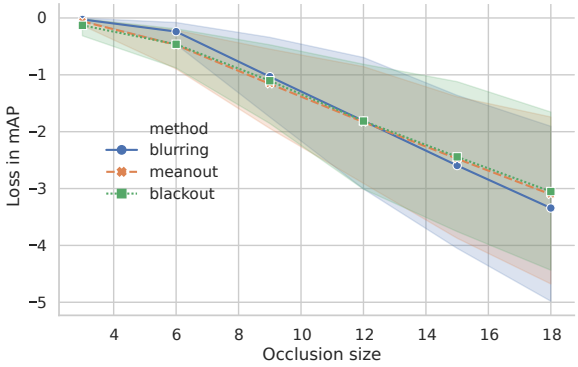


Fig. 4: The relation between occlusion size and average loss in performance for keypoint level methods. Occlusion size greatly affects the performance.

with left and right ankle affected the most. Interestingly, attacking on one side improves performance of the the other side (Figure 8c). Attacking on left side increases the mAP score of right side such as shoulder, ear, elbow keypoints. The analysis demonstrates that the pose estimator is sensitive to part occlusion attacks.

IV. Occlusion Augmentation Against Attacks

We evaluate two main human pose estimation datasets: COCO [26] has 200k images with 250k person instances, labelled with 17 keypoints and MPII [1] has 40k persons, each labelled with 16 joints. The train, validation and test sets include 22k, 3k and 15k person instances respectively. For the evaluation of MPII dataset, the validation set is used since the labels of the test set are not available.

For training HRNet [42] models on COCO [26] and MPII [1] we follow the original pipeline of HRNet. For COCO dataset, human detection boxes are extended to fit 4:3 aspect ratio, and cropped from the image and resized to $256x192$. The pose estimator is trained with the keypoint location of the joints. The data augmentations that are used in HRNet training include random rotation $\in [-45^o, 45^o]$, random scale $\in [0.65, 1.35]$, random flipping and half-body augmentations. The Adam optimizer [20] is used to train the network with the learning rate schedule following [53], starting with $1e-3$ and reduced to $1e-4$ and $1e-5$ at 170th and 200th epochs respectively and the training is completed at the 210th epoch. For MPII dataset, the training procedure of HRNet is as followed: 256x256 input size is used and half-body augmentations are discarded. For the evaluation of the models, Object Keypoint Similarity (OKS) for COCO and Percentage of Correct Keypoints (PCK) for MPII are used.

During testing, HRNet firstly employs an object detection algorithm to obtain boxes with a single person. Afterwards the pose estimator produces the keypoint location of the joints.

A. Occlusion augmentation

We investigate the following three methods: (i) Targeted Blurring, (ii) Targeted Cutout, (iii) Targeted PartMix. The augmentation techniques are called as *targeted*, because we

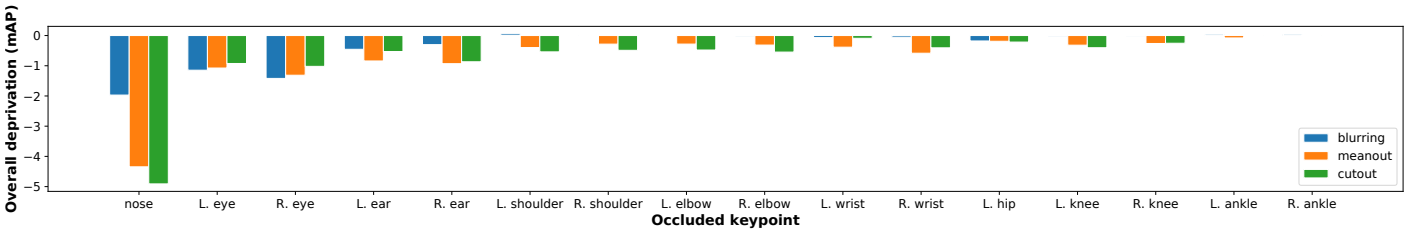
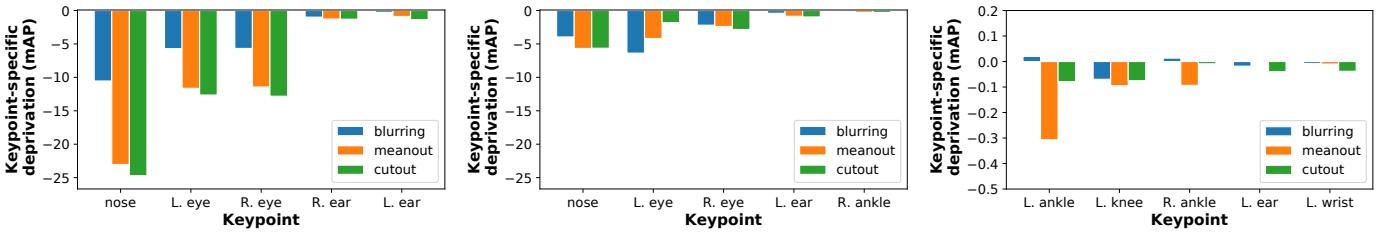


Fig. 5: Overall loss in mAP after performing keypoint level occlusion. *L.* and *R.* correspond to the left and right side respectively. To note that, the occluded keypoint is included in the evaluation. Occluding nose causes the highest loss in performance.



(a) The nose is the most influential keypoint causes a significant drop in the performance for the closest keypoints - left eye and right eye by around 10%.

(b) When we occlude the left eye, there is a smaller loss in keypoint-specific performance for the left eye than while occluding nose.

(c) Left ankle is one of the least influential keypoints with loss only visible for meanout for occluded keypoint.

Fig. 6: Loss in AP for top 5 keypoints with largest deprivation, when an individual key point is occluded.

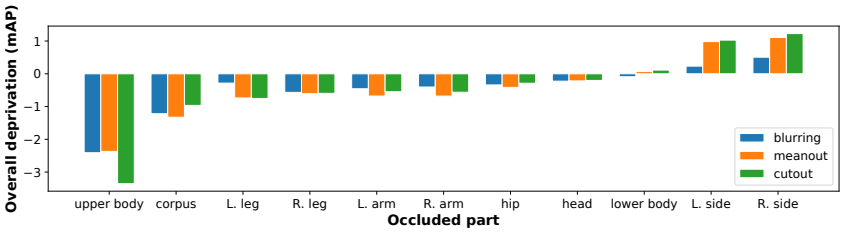
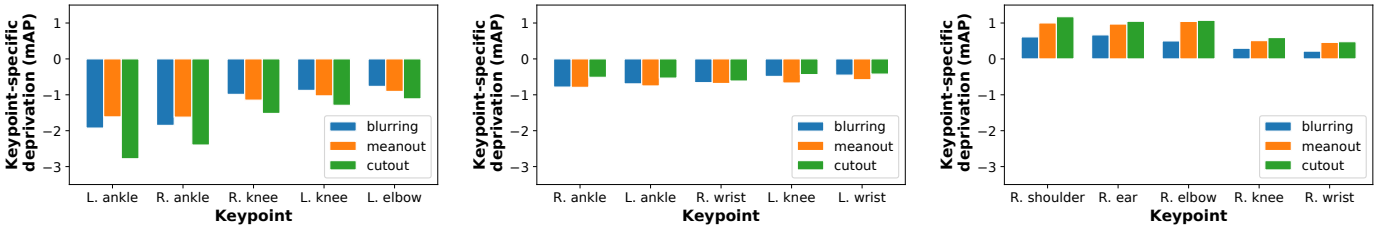


Fig. 7: Change in mAP for various parts occluded. Upper body and sides are the parts that cause the highest loss in the performance.



(a) Significant loss in performance for all of the remaining keypoints. Blackout affects the method most.

(b) Similar loss across remaining keypoints, indicating that corpus is one of the most influential parts.

(c) Occluding the left side of the body improves the performance of right shoulder, ear and elbow.

Fig. 8: Change in AP for top 5 keypoints with the largest difference, when chosen part is occluded.

apply them on target locations of keypoints or parts instead of random location in the image. It is important to state that the proposed augmentation techniques are introduced after the bounding box person detection, and it thus does not affect the object detection method.

Targeted Blurring. We use Gaussian blur for two types of targeted blurring: (i) keypoint blurring with a kernel size of 9 pixels (Figure 9a) and (ii) part blurring with a kernel size of 31 pixels shown in Figure 9d.

Targeted Cutout. The size of the keypoint cutout (Figure 9b-9c) and part cutout (Figure 9e) are similar to the blurring equivalents. Instead of blurring, the area is colored with mean value of the image.

Targeted PartMix. The method is designed to mitigate

(a) Blurring on left hip (almost not visible).

(b) Cutout on left hip.

(c) Multi keypoint Cutout.

(d) Part Blurring.

(e) Part Cutout.

(f) PartMix on the right leg with a pasted random arm.

Fig. 9: Targeted keypoint augmentations: a, b, c and targeted part augmentations: d, e, f.

the occlusions caused by another person (Figure 9f). In this approach, a different part from a random image is pasted in the place of a body part area. In this process, the keypoint labels of newly pasted part are not introduced to heatmap labels. This augmentation is only performed on body parts. Similar to the part level blurring and cutout augmentation methods, the occluded keypoints under the pasted area are still predicted.

B. Analyses of occlusion augmentation

All the following augmentation methods, except baselines, already include flipping, rotation, scaling and half-body augmentations. Each network obtains the boxes from Cascade RCNN [2] detector which has ResNet50 backbone. The results of each method can be seen in Table I.

Baselines. Table I indicates 3 baseline variants. Firstly, HRNet without any augmentations obtains only 65.3% mAP score. Secondly, adding flipping, rotation and scaling augmentations improve non-augmented baseline by 8.6%. Last variant is half body augmentation which adds only 0.4% improvements on rotation and scaling augmentations.

Single keypoint augmentations. We check the performance of 3 different augmentations: blurring, cutout and a combination of two of them which are applied on a single keypoint with the varying probability of 0.2 and 0.5 (Figure 9a-9b). We observe the highest improvement for blurring and cutout by 0.2% when the probability is chosen as 0.5 (Table I). Other single keypoint variants do not improve the performance.

Multi-keypoint augmentations. We applied random multi-keypoint variant blurring and cutout with a maximum of 5 keypoints with a probability of 0.2 (Figure 9c). The augmentation decreases the model performance by 0.4%.

Part augmentations. 4 different part augmentation methods are used: part blurring, part cutout, a combination of both them and PartMix (Figure 9d, 9e and 9f respectively). To demonstrate the effect of each augmentation, we apply them with a probability of 0.2 and 0.5. In addition, the effect of removing the labels of the occluded keypoint is also investigated as *removal* column in Table I.

In the bottom part of Table I, cutout and PartMix show 0.2% and 0.1% improvements respectively. In all the variants of blurring, small degradation or no improvement is observed. The combination of part level variants of cutout and blurring indicate some decreases of the performance for the removal configuration with probability of 0.2 and 0.5 and do not improve in non-removal configuration.

To conclude to findings from the Table I, flipping, rotation and scaling augmentations add a huge performance gain to the HRNet. However, including half-body, the occlusion based augmentation methods do not improve the performance of the pose estimator significantly.

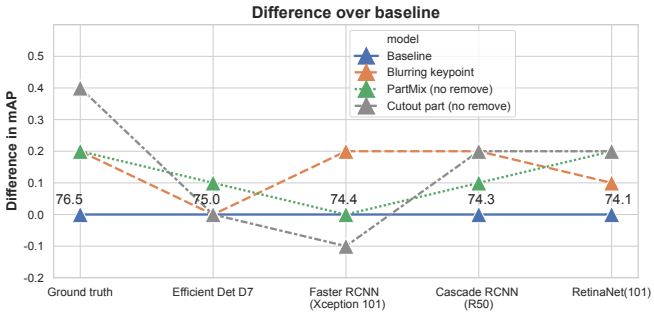


Fig. 10: Performance of chosen augmentations for HRNet-32 on various detection backbones and ground truth boxes. The ground truth bounding box performs best. Yet, none of the data augmentation methods help to improve performance over 0.2% for any object detector.

The effect of the object detection algorithms. HRNet [42] is a top-down approach which utilizes an object detection algorithm to obtain human instances. Therefore, the performance of the pose estimation considerably depends on the detection performance, namely detected human instances.

By the evidence of the Table I, we choose keypoint blurring, part cutout and PartMix methods for further analysis as they are the most promising augmentations.

We evaluate the pose estimation performances of vanilla HRNet and also of HRNet with the chosen augmentation methods with two 2-stage detectors, Faster RCNN [35] with XCeption 101 backbone and Cascade RCNN [2]; 2 single-stage detectors, RetinaNet [27] and EfficientDet D7 [43]; and by using ground truth boxes of human instances (Figure 10).

Augmentation	*level*	*removal*	*p*	**Evaluation results** AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Baseline (no augments)	-	-	-	65.3	86.4	72.6	62.6	70.7	70.2
Baseline (flip, rot, scale)	-	-	-	73.9	90.0	80.9	70.4	80.3	78.3
Baseline (flip, rot, scale, half-body)	-	-	-	74.3	90.6	81.7	70.7	80.7	78.8
Blurring	k	✗	0.2	74.3	90.4	81.6	70.8	80.6	78.7
	k	✗	0.5	74.5	90.4	81.8	70.8	80.8	78.7
Cutout	k	✗	0.2	74.3	90.4	81.7	71.0	80.3	78.7
	k	✗	0.5	74.5	90.5	81.7	70.9	80.7	78.8
Cutout + Blurring	k	✗	0.2	74.0	90.4	81.1	70.4	80.3	78.4
	k	✗	0.5	74.3	90.5	81.1	70.8	80.6	78.6
Blurring	p	✓	0.2	74.3	90.5	81.7	70.6	80.8	78.6
	p	✓	0.5	74.0	90.5	81.1	70.5	80.4	78.4
	p	✗	0.5	74.1	90.3	81.1	70.6	80.2	78.5
Cutout	p	✓	0.2	74.2	90.5	81.2	70.8	80.4	78.6
	p	✓	0.5	74.2	90.3	81.1	70.6	80.4	78.6
	p	✗	0.5	74.5	90.5	81.6	70.9	80.7	78.8
Cutout + Blurring	p	✓	0.2	73.4	90.3	80.8	69.9	79.5	77.8
	p	✓	0.5	73.9	90.4	81.0	70.5	80.0	78.3
	p	✗	0.5	74.3	90.4	81.2	70.6	80.5	78.6
Multikeypoint (max. 5)	-	-	0.2	73.9	90.1	80.9	70.5	80.2	78.3
PartMix	-	✓	0.5	74.3	90.5	81.1	70.7	80.6	78.7
	-	✗	0.5	74.4	90.7	81.5	71.1	80.5	78.8

TABLE I: Comparison of augmentation variants on COCO validation set for HRNet using CascadeRCNN bounding boxes. Upper-part indicates single-keypoint augmentation and bottom-part shows multiple-keypoint augmentation. k and p in the level column represent keypoint and part augmentations respectively. Removal column indicates if the occluded keypoints are removed from prediction. Column p is the probability of augmentation. Keypoint cutout and blurring, and part cutout and PartMix improve the performance. Other variants obtain results either on a par with baseline or worse than baseline.

Augmentation	*level*	*remove*	*p*	**Evaluation results** *Head*	*Shoulder*	*Elbow*	*Wrist*	*Hip*	*Knee*	*Ankle*	***Total***
Baseline	-	-	-	97.1	95.9	90.4	86.4	89.1	87.2	83.3	90.3
Blurring	k	✗	0.5	97.3	95.9	90.5	86.2	89.2	86.4	83.1	90.3
Cutout	p	✗	0.5	97.2	96.3	90.7	86.7	89.4	86.7	83.3	**90.5**
PartMix	-	✗	0.5	97.4	96.2	91.0	86.8	89.2	86.7	83.0	**90.5**

TABLE II: Results on MPII dataset. Keypoint blurring obtains on a par with the HRNet baseline, yet part cutout and PartMix increase the performance.

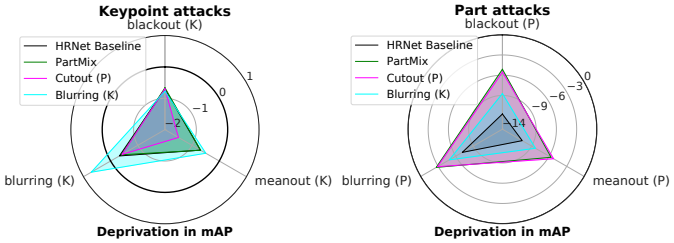
All the augmentations indicate improvements using ground truth bounding boxes by 0.2% for keypoint blurring and PartMix, and 0.4% for part cutout. All the chosen augmentation methods obtain better result with Cascade RCNN and RetinaNet $0.1 - 0.2\%$ depending on the augmentation. With EfficientDet D7 detector, keypoint blurring and part cutout result in similar to baseline except 0.1% improvement by PartMix. For Faster-RCNN, keypoint blurring shows 0.2% increase, yet part cutout degrades the performance by 0.1%.

The performances of baseline and the augmentations vary depending on the object detector. The augmentation methods improves the results slightly, yet the gain is insignificant.

Performance on MPII. We also test the data augmentation methods on MPII dataset (Table II). If we check the total contribution of the proposed augmentations, keypoint blurring result in on a par with baseline, yet part cutout and PartMix increase the performance by 0.2% for the metric PCK@0.5. The largest improvement per keypoint is observed for elbows by 0.6% and wrists by 0.3%, with the degradation on knees and ankles by 0.4% and 0.2% respectively.

Similar to analyses on the COCO dataset, the proposed augmentations can only improve the performance slightly.

How occlusion robust is data augmentation? Figure 11 shows the robustness of the baseline and the proposed augmentations to the occlusion attacks. The analysis is done on COCO dataset and the results are shown as mAP score of all keypoints. We can clearly see that training with the keypoint blurring augmentation makes the network more robust against blurring attack, but there is no significant improvement for the other keypoint attacks. In case of part attacks, we observe an improvement across all augmentation methods over the baseline. For the part augmentations, there is a significant improvement against all part level attacks in comparison to baseline. Specifically, PartMix has almost no advantages



(a) (b)

Fig. 11: Robustness comparison of proposed methods against (a) keypoint and (b) part occlusion attacks. Part augmentations improve the baseline but does not solve occlusion.

	Evaluation results				
Augmentation	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Higher HRNet	67.1	86.2	73.0	61.5	76.1
Blurring (K)	66.5	86.3	72.1	60.6	75.7
Cutout part (no remove)	66.6	86.4	72.9	60.7	75.6
PartMix (no remove)	67.0	86.4	73.0	61.3	75.8

TABLE III: Results for bottom-up method, Higher HRNet [6]. The keypoint blurring, part cutout and Partmix degrade the performance of bottom-up methods. The augmentations do not help Higher HRNet.

against keypoint attacks, however, it improves part level methods about more than 5% in comparison to baseline. Part cutout obtains similar performance with PartMix against part attacks. Proposed augmentations reduce the performance deprivations when we apply occlusion attacks, yet data augmentation still does not solve the occlusion problem.

C. Augmentation on bottom-up method: Higher HRNet

We also apply occlusion augmentations on Higher HRNet [6], a bottom-up method. Higher HRNet is built on HRNet-32 and inputs 512x512 sized images. The training procedure follows Higher HRNet implementation from the paper. Unlike top-down methods, Higher HRNet operates on full-image and try to obtain the keypoints of each instance from the full-image. When applying the augmentations on Higher HRNet, we target all the human instances in the image.

Results in Table III show the augmentation methods to improve AP50 score slightly. For AP, all augmentations degrade performance by 0.6% for keypoint level blurring, by 0.5% for part level cutout and by 0.1% for PartMix. Hence, using part and keypoint augmentations do not improve the performance of a bottom-up method.

V. Discussion and Conclusion

In this study, we investigate the sensitivity of human pose estimators to occlusion. Firstly, we introduce targeted keypoint and body part occlusion attacks to show how much occlusion affects the performance. Secondly, keypoint and part based data augmentation techniques against occlusion are investigated. The structured analyses indicate that deep pose estimators are not robust to occlusion. With all the bells and whistles, the current and proposed data augmentation methods do **not** bring significant improvements on the performance of the top-down pose estimators and even reduce the performance for the bottom-up approaches. Our paper is important because it helps data scientists looking for improvements against occlusions to not work on data augmentation. Battling occlusions is still an open problem for human pose estimation.

Part based attacks and augmentation are applied as a rectangle shape. This fact can introduce unusual artefacts because natural occlusions can have arbitrary shapes. Each keypoint augmentation is applied as a circle that covers the related keypoint, yet in reality, keypoint occlusions can occur with numerous shapes and ways e.g. self occlusion, occlusion by other object. Moreover, for bottom-up approaches, the input image into the network may have more perturbations since the full image can contain multiple instances. This fact can harm the learning process.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1, 3

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018. 5

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1812.08008, 2018. 1, 2

[4] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4733–4742, 2016. 2

[5] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2

[6] 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 *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 7

[7] J. Choe, S. Lee, and H. Shim. Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 1, 2

[8] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. pages 5669–5678, 07 2017. 2

[9] Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017. 1, 2

[10] Philipe Ambrozio Dias, Damiano Malafronte, Henry Medeiros, and Francesca Odone. Gaze estimation for assisted living environments. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 290–299, 2020. 1

[11] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 364–380, 2018. 1, 2

[12] Zhijie Fang and Antonio M. López. Intention recognition of pedestrians and cyclists by 2d pose estimation. *ArXiv*, abs/1910.03858, 2019. 1

[13] Geoffrey French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham D. Finlayson. Consistency regularization and cutmix for semi-supervised semantic segmentation. *CoRR*, abs/1906.01916, 2019. 1, 2

[14] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. Detect-and-Track: Efficient Pose Estimation in Videos. In *CVPR*, 2018. 1

[15] Thomas Golda, Tobias Kalb, Arne Schumann, and Jürgen Beyerer. Human pose estimation for real-world crowded scenarios. *CoRR*, abs/1907.06922, 2019. 1

[16] Hongyu Guo, Yongyi Mao, and Richong Zhang. Mixup as locally linear out-of-manifold regularization. *CoRR*, abs/1809.02499, 2018. 1, 2

[17] Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[18] Ying Huang, Bin Sun, Haipeng Kan, Jiankai Zhuang, and Zengchang Qin. Followmeup sports: New benchmark for 2d human keypoint recognition. *ArXiv*, abs/1911.08344, 2019. 1

[19] Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. Multi-scale structure-aware network for human pose estimation. *CoRR*, abs/1803.09894, 2018. 1, 2

[20] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014. 3

[21] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. *CoRR*, abs/1903.06593, 2019. 1, 2

[22] L. Ladický, P. H. S. Torr, and A. Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3578–3585, 2013. 1

[23] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *ArXiv*, abs/1901.00148, 2019. 2

[24] Zhong Li, Xin Chen, Wangyiteng Zhou, Yingliang Zhang, and Jingyi Yu. Pose2body: Pose-guided human parts segmentation. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 640–645. IEEE, 2019. 1

[25] Kevin Lin, Lijuan Wang, Kun Luo, Yinpeng Chen, Zicheng Liu, and Ming-Ting Sun. Cross-domain complementary learning using pose for multi-person part segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. 1

[26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 1, 2, 3

[27] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 5

[28] Diogo Luvizon, David Picard, and Hedi Tabia. Multi-task deep learning for real-time 3d human pose estimation and action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2020. 1

[29] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 1, 2

[30] Xiaohan Nie, Caiming Xiong, and Song-Chun Zhu. Joint action recognition and pose estimation from video. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1293–1301, 2015. 1

[31] Štěpán Obdržálek, Gregorij Kurillo, Ferda Ofli, Ruzena Bajcsy, Edmund Seto, Holly Jimison, and Michael Pavel. Accuracy and robustness of kinect pose estimation in the context of coaching of elderly population. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1188–1193. IEEE, 2012. 1

[32] Sai Perla, Sudip Das, Partha Mukherjee, and Ujjwal Bhattacharya. Cluenet : A deep framework for occluded pedestrian pose estimation. 12 2019. 1, 2

[33] U. Rafi, J. Gall, and B. Leibe. A semantic occlusion model for human pose estimation from a single depth image. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 67–74, 2015. 1, 2

[34] N. D. Reddy, M. Vo, and S. G. Narasimhan. Occlusion-net: 2d/3d occluded keypoint localization using graph networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7318–7327, 2019. 2

[35] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015. 5

[36] Sahil Shah, Naman Jain, Abhishek Sharma, and Arjun Jain. On the robustness of human pose estimation, 08 2019. 1, 2

[37] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019. 1

[38] K. K. Singh and Y. J. Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3544–3553, 2017. 1, 2

[39] Anna Sokolova and Anton Konushin. Pose-based deep gait recognition. *CoRR*, abs/1710.06512, 2017. 1

[40] Khurram Soomro, Haroon Idrees, and Mubarak Shah. Online localization and prediction of actions and interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:459–472, 2019. 1

[41] C. Summers and M. J. Dinneen. Improved mixed-example data augmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1262–1270, 2019. 1

[42] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep High-Resolution Representation Learning for Human Pose Estimation. 2019. 1, 2, 3, 5

[43] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. *ArXiv*, abs/1911.09070, 2019. 5

[44] H. L. Tavares, J. B. C. Neto, J. P. Papa, D. Colombo, and A. N. Marana. Tracking and re-identification of people using soft-biometrics. In *2019 XV Workshop de Visão Computacional (WVC)*, pages 78–83, 2019. 1

[45] Luke Taylor and Geoff Nitschke. Improving deep learning using generic data augmentation. *arXiv preprint arXiv:1708.06020*, 2017. 1

[46] Y. Tokozume, Y. Ushiku, and T. Harada. Between-class learning for image classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5486–5494, 2018. 1, 2

[47] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, 2014. 2

[48] Jagannadan Varadarajan, Ramanathan Subramanian, Samuel Rota Bulò, Narendra Ahuja, Oswald Lanz, and Elisa Ricci. Joint estimation of human pose and conversational groups from social scenes. *International Journal of Computer Vision*, 126(2-4):410–429, 2018. 1

[49] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6438–6447, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 1

[50] Sijia Wang, Fabian Flohr, Hui Xiong, Tuopu Wen, Bao feng Wang, Mengmeng Yang, and Diange Yang. Leverage of limb detection in pose estimation for vulnerable road users. *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 528–534, 2019. 1

[51] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016. 2

[52] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L. Yuille. Joint multi-person pose estimation and semantic part segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[53] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. *CoRR*, abs/1804.06208, 2018. 1, 2, 3, 9

[54] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *BMVC*, 2018. 1

[55] Huayong Xu, Yangyan Li, Wenzheng Chen, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. A holistic approach for data-driven object cutout. *CoRR*, abs/1608.05180, 2016. 1, 2

[56] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. *CoRR*, abs/1708.01101, 2017. 2

[57] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *CoRR*, abs/1905.04899, 2019. 2

[58] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. *ArXiv*, abs/1910.06278, 2019. 2

[59] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412, 2017. 1, 2

[60] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *CoRR*, abs/1708.04896, 2017. 1, 2

APPENDIX

MORE RESULTS ON COCO VAL SET

HRNet results. For this experiment, we increase the input resolution of images from 256x192 to 384x256. The training process follows the aforementioned scheme for COCO dataset.

According to the analysis of the performance across a variety of detection backbones shown in Figure 12, we notice that PartMix is consistently improving performance - with the greatest boost of 0.4% for Cascade R-CNN and 0.3% for Faster RCNN. For both keypoint blurring and part cutout, we observe no significant improvement or even the performance decreases - for part cutout using EfficientDet, Faster RCNN and RetinaNet and for Blurring using RetinaNet. All the presented augmentations show largest gain for Cascade RCNN. Occlusion augmentations do not help to solve occlusion problems when higher resolution input is used.

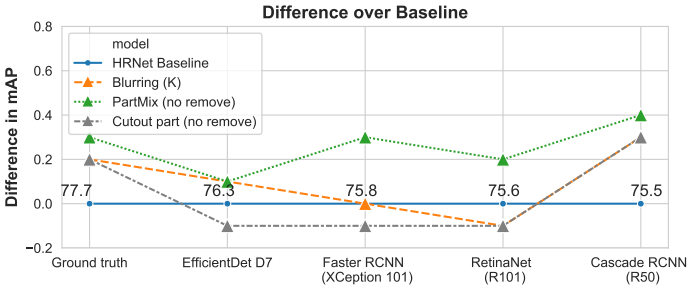


Fig. 12: Higher resolution input for HRNet 32: the resolution is changed from 256x192 to 384x256. The best performance across detection backbones is observed for PartMix.

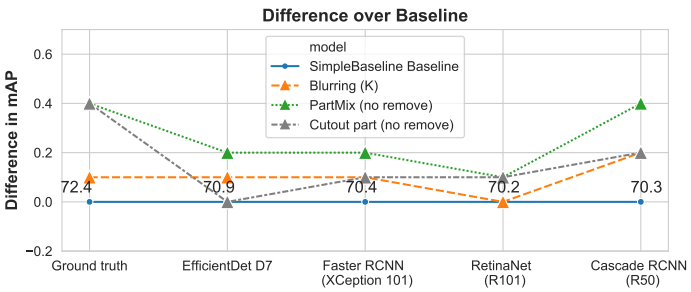


Fig. 13: Performance of chosen augmentations for SimpleBaseline on various detection backbones and ground truth boxes. Using the ground truth bounding boxes outperforms all the SimpleBaseline methods with a detection backbones.

SimpleBaseline results. The usability of occlusion augmentations are not only limited to HRNet, yet they can be used with other top-down methods like SimpleBaseline [53]. In this experiment, we apply the occlusion augmentations on SimpleBaseline method with different object detection backbones. The training procedure of the network follows the original implementation.

By checking the performance across the various detection backbones we observe either small or no improvement at all (Figure 13). PartMix show the most significant improvement across detection backbones, with 0.4% boost in the performance for the ground truth boxes and the boxes produced by Cascade RCNN, 0.2% for EfficientDet and Faster RCNN and 0.1 % for RetinaNet. Cutout and Blurring improve at most 0.2% across all the detection backbones, apart from 0.4% for Cutout using ground truth bounding boxes. According to the results, proposed augmentation techniques do not solve occlusion problems of SimpleBaseline method.

Fig. 14: Qualitative comparison between ground truth (left), baseline (middle) and keypoint Blurring (K) (right). 1st and 2nd rows respectively - misplacement of left wrist keypoint and mismatch between knee keypoints in the baseline and keypoint blurring fixes the mistakes. 3rd row - both baseline and proposed method produce wrong keypoints. 4th row - baseline produces near-optimal keypoints whilst keypoint blurring makes mistake on left ankle keypoint. Data augmentation does not solve occlusion problem.

VISUALIZATION OF RESULTS

Figure 14 presents a qualitative comparison between ground truth, HRNet-32 Baseline and keypoint blurring augmentation. In the first and second rows, keypoint blurring outperforms the baseline by obtaining the position of the left wrist and knee keypoints respectively. In the third row, both baseline and keypoint blurring produce wrong keypoint predictions. Fourth row presents failure case when baseline produces near-optimal annotations, while the method with keypoint blurring predicts left ankle in place of the right one.