Joint COCO and Mapillary Workshop at ICCV 2019: COCO Keypoint Detection Task Challenge Track

Technical Report: PifPaf

Sven Kreiss Alexandre Alahi
VITA lab at EPFL
sven.kreiss@epfl.ch

Abstract

We are building on our CVPR 2019 paper [6] and extending it with new augmentations and a strategy for multi-scale fusion at the field level. We show qualitative results of the intermediate field representations, ablation studies and timing numbers. For easy comparison, we show our numbers for single- and multi-scale predictions on both the validation and test-dev sets. We set new state-of-the-art results for a bottom-up method that does not use additional data. The code is available at github.com/vita-epfl/openpifpaf.

1. Method

Augmentations. The main change for our single-scale performance boost are augmentations. We increased the range of the random re-scaling from $r \in [0.5, 1.0]$ to $r \in [0.5, 2.0]$. We do not sample uniformly from this range as this would produce twice as many upsampled images than downsampled. Instead we compute

$$r = 2^u \quad \text{where} \quad u \sim \mathcal{U}(-1, 1)$$

where u is drawn from the uniform distribution $\mathcal{U}$ over the symmetric interval $[-1, 1]$ and which generates reduced resolutions ($r < 1$) for negative u and increased resolutions ($r > 1$) for positive u. As this effectively doubled the variation in input images, we also doubled the training epochs from 75 to 150. Our implementation generalizes to arbitrary lower and upper boundaries $r \in [l, h]$ and draws $u \sim \mathcal{U}(\log_2 l, \log_2 h)$.

We also adjusted random cropping to crop randomly to an area-of-interest where area-of-interest is the image region with person annotations plus a large 50 px margin.

Reduced jitter. We apply a small modification to the algorithm that selects the connection from one joint to the next. Similar to [1], instead of selecting the best connection, we select the best two connections and connect to the weighted average of these two connections. However, this is only done if the second best connection has a confidence that is at least half as large as the best connection.

Instance score. The instance score is the mean of the keypoint scores where the three highest keypoint scores are weighted three times higher.

Multi-scale field fusion. Multi-scale outputs can be fused at various levels: at the feature map level, field level and at the level of reconstructed poses. Fusing at feature map level can erase information and fusing at the level of reconstructed poses does not share information across resolutions. To profit optimally from our architecture, we choose to fuse at field level.

Our backbone processes 10 images with resolution factors $\{1.5, 1, 0.75, 0.5, 0.3\}$ and the same again horizontally flipped. To un-flip the output fields, the following operations are performed: reverse the x-locations, invert the x-coordinate of all vector components, switch left and right fields and for PAF fields switch the direction of some fields, *e.g.*, the connection between left and right hip is still the same field but in the opposite direction.

Processing 10 images independently is less efficient than single-network multi-resolution networks, but it allows us to be backbone agnostic. The PifPaf head networks are optimized for low resolution and so the input resolution required by PifPaf is lower than for many other methods leading also to efficient processing.

We use the same high resolution aggregation map for PIF vectors as in the single-scale setting. In the multi-scale setting, we aggregate within one scale together with the horizontally flipped version to produce five maps. The five maps are merged into a single map with a max operation.

PAF fields are scaled to a common resolution. PAF fields are never stored in rasterized form. The lists of fields are concatenated. To compartmentalize the resolutions, minimum and maximum distance thresholds are applied, *e.g.*,

Figure 1: Seeds color-coded by the joint-type.

	AP	AP^M	AP^L
OpenPose [2]	61.8	57.1	68.2
Associative Embedding [7]	65.5	60.6	72.6
PersonLab [8] – single-scale	66.5	62.4	72.3
PifPaf – single-scale (ours)	68.2	63.5	74.8
PersonLab [8]	68.7	64.1	75.5
MultiPoseNet [5]	69.6	65.0	76.3
HigherHRNet [3]	70.5	66.6	75.8
PifPaf (ours)	**71.2**	**67.4**	**76.9**

Table 1: Metrics in percent evaluated on the COCO 2017 test-dev for bottom-up methods.

	AP [%]	t [ms]	t^{dec} [ms]
our CVPR2019 [6]	67.4	263	173
single-scale w/o reduced jitter	69.7	151	55
single-scale	69.8	153	56
multi-scale w/o hflip	72.2	853	317
multi-scale	73.0	1507	433

Table 2: Interplay between precision and single-image prediction time t of ResNet152 models on a GTX1080Ti for the COCO val set. Last column is the decoding time t^{dec}.

the highest resolution is responsible to make connections of distances from 0.0 to 160 px.

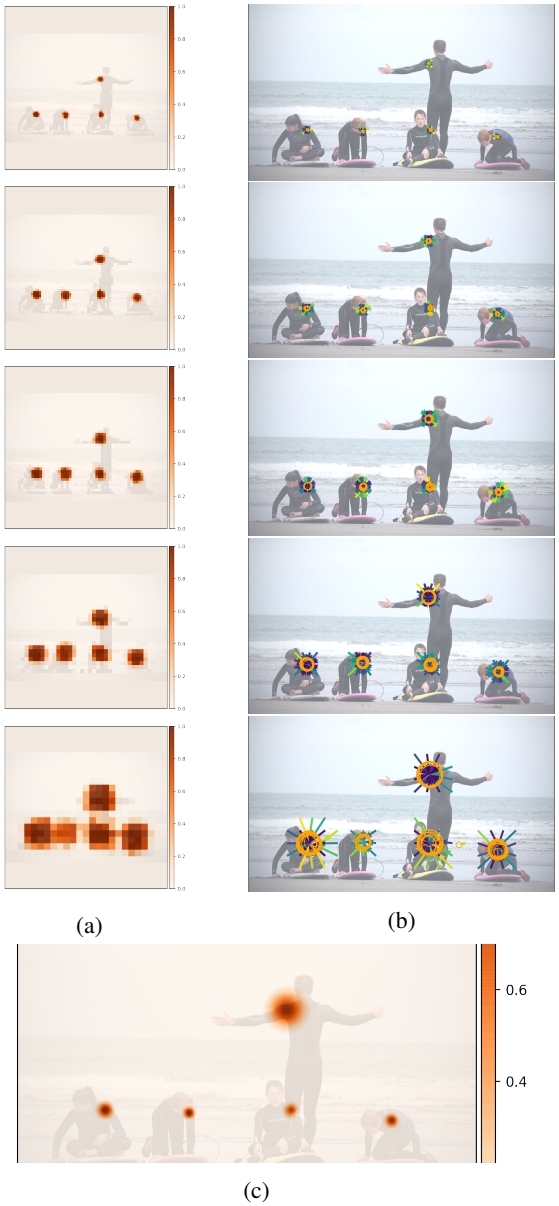
2. Experiments

We use ImageNet [4] pretrained models and then refine on COCO keypoint annotations. We do not use any additional datasets, in particular no additional pose annotations.

Qualitative representations of the seeds from which poses are started are shown in Figure 1. PIF fields are shown in Figure 2 and PAF fields in Figure 3.

Main result on the test-dev set is shown in Table 1. On the test-challenge dataset, we reach 69.4% in overall AP and 64.7% and 76.5% for medium and large object AP.

Precision and timing results are shown in Table 2. All

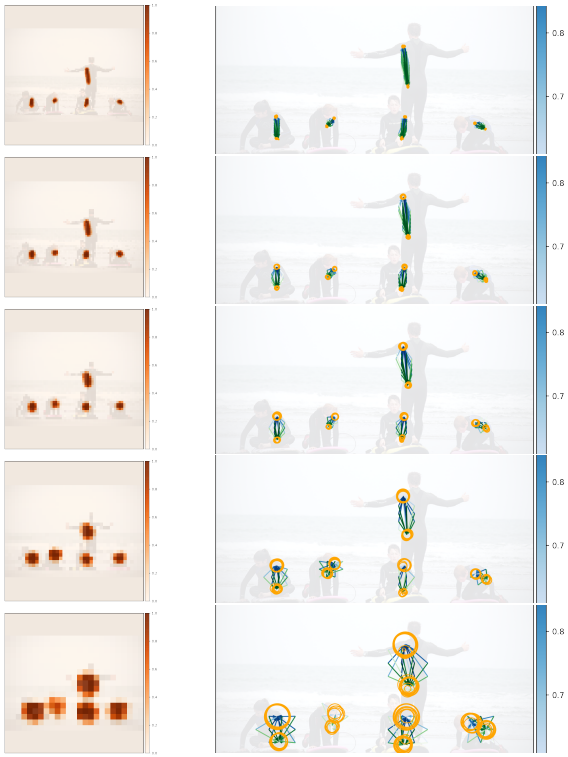


(a) (b)

(c)

Figure 2: Multi-scale PIF components for the left shoulder: (a) shows the confidence component and (b) the vector component. The magnitude of the spatial uncertainty b of each regression is indicated by an orange circle. All resolutions are aggregated into a single high resolution map shown in (c).

results are produced with images where the longer edge is scaled to 641 px or where this is used as the basis length for multiple scales.



(a) (b)

Figure 3: Multi-scale PAF components for the left shoulder-hip connection: (a) shows the confidence component and (b) the vector components. The magnitude of the spatial uncertainty b of each regression is indicated by an orange circle.

References

[1] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047*, 2019. 1

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, volume 1, page 7, 2017. 2

[3] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Bottom-up higher-resolution networks for multi-person pose estimation. *arXiv preprint arXiv:1908.10357*, 2019. 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[5] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 417–433, 2018. 2

[6] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2

[7] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems*, pages 2277–2287, 2017. 2

[8] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. *CoRR*, abs/1803.08225, 2018. 2