A. Appendix

A.1. Generalized Coordinate Transformation

In Sec. 3.4 we have assumed $\sigma_{\text{HW}}{=}\hat{\sigma}_{\text{HW}}$ and $\sigma_{\text{VU}}{=}\hat{\sigma}_{\text{VU}}$. Here we relax this condition and only assume $\sigma_{\text{HW}}{=}\hat{\sigma}_{\text{HW}}$. Again, we still have the following two relations for x, u: $x{+}\alpha u{=}\hat{x}$ and $x{=}\hat{x}{-}\hat{\alpha}\hat{u}$. Solving for $\hat{x}$ and $\hat{u}$ gives: $\hat{x}{=}x{+}\alpha u$ and $\hat{u}{=}\frac{\alpha}{\hat{\alpha}}u$. Then `align2nat` is:

$$\mathcal{F}(v,u,y,x) = \hat{\mathcal{F}}(\frac{\alpha}{\hat{\alpha}}v, \frac{\alpha}{\hat{\alpha}}u, y{+}\alpha v, x{+}\alpha u). \tag{2}$$

More generally, consider arbitrary units σ_{HW}, $\hat{\sigma}_{\text{HW}}$, σ_{VU}, and $\hat{\sigma}_{\text{VU}}$. Then the relations between the natural and aligned representation can be rewritten as:

$$\begin{cases} x{\cdot}\sigma_{\text{HW}}{+}u{\cdot}\sigma_{\text{VU}} &= \hat{x}{\cdot}\hat{\sigma}_{\text{HW}} \\ x{\cdot}\sigma_{\text{HW}} &= \hat{x}{\cdot}\hat{\sigma}_{\text{HW}}{-}\hat{u}{\cdot}\hat{\sigma}_{\text{VU}} \end{cases} \tag{3}$$

Note that these relations only hold in the image pixel domain (hence the usage of all units). Solving for $\hat{x}$, $\hat{u}$ gives:

$$\begin{cases} \hat{x} &= \frac{\sigma_{\text{HW}}}{\hat{\sigma}_{\text{HW}}}x{+}\frac{\sigma_{\text{VU}}}{\hat{\sigma}_{\text{HW}}}u \\ \hat{u} &= \frac{\sigma_{\text{VU}}}{\hat{\sigma}_{\text{VU}}}u \end{cases} \tag{4}$$

And the `align2nat` transform becomes:

$$\mathcal{F}(v,u,y,x) = \hat{\mathcal{F}}(\frac{\sigma_{\text{VU}}}{\hat{\sigma}_{\text{VU}}}v, \frac{\sigma_{\text{VU}}}{\hat{\sigma}_{\text{VU}}}u, \frac{\sigma_{\text{HW}}}{\hat{\sigma}_{\text{HW}}}y{+}\frac{\sigma_{\text{VU}}}{\hat{\sigma}_{\text{HW}}}v, \frac{\sigma_{\text{HW}}}{\hat{\sigma}_{\text{HW}}}x{+}\frac{\sigma_{\text{VU}}}{\hat{\sigma}_{\text{HW}}}u). \tag{5}$$

This version of the coordinate transformation demonstrates the role of units and may enable more general uses.

A.2. Aligned Representation and InstanceFCN

We prove that the InstanceFCN [7] output behaves as an upscaling aligned head with *nearest-neighbor* interpolation.

In [7], each output mask has $V{\times}U$ pixels that are divided into $K{\times}K$ bins. A mask pixel is read from the channel corresponding to the pixel's bin. In our notation, [7] predicts $\mathcal{G}$ which is related to the natural representation $\mathcal{F}$ by:

$$\mathcal{F}(v,u,y,x) = \mathcal{G}([\frac{K}{V}v], [\frac{K}{U}u], y+v, x+u), \tag{6}$$

where $[\cdot]$ is a rounding operation and the integers $[\frac{K}{V}v]$ and $[\frac{K}{U}u]$ index a bin. Now, define a new function $\tilde{\mathcal{F}}$ by:

$$\tilde{\mathcal{F}}(v,u,y{+}v,x{+}u) \triangleq \mathcal{G}([\frac{K}{V}v], [\frac{K}{U}u], y{+}v, x{+}u), \tag{7}$$

and new coordinates: $\tilde{x}{=}x{+}u$ and $\tilde{u}{=}u$ (likewise for v and y). Then $\tilde{\mathcal{F}}$ can be written as:

$$\tilde{\mathcal{F}}(\tilde{v},\tilde{u},\tilde{y},\tilde{x}) \triangleq \mathcal{G}([\frac{K}{V}\tilde{v}], [\frac{K}{U}\tilde{u}], \tilde{y}, \tilde{x}). \tag{8}$$

Eqn.(8) says that $\tilde{\mathcal{F}}$ is the *nearest-neighbor* interpolation of $\mathcal{G}$ on $(\tilde{V}, \tilde{U})$. Eqn.(7), (6), and the new coordinates show that $\mathcal{F}$ is computed from $\tilde{\mathcal{F}}$ by the `align2nat` transform with $\alpha{=}1$. Thus, InstanceFCN masks can be constructed in the TensorMask framework by predicting $\mathcal{G}$, performing nearest-neighbor interpolation of $\mathcal{G}$ on $(\tilde{V}, \tilde{U})$ to get $\tilde{\mathcal{F}}$, and then using `align2nat` to compute natural masks $\mathcal{F}$.

A.3. Object Detection Results

In Tab. 4 we show the associated *bounding-box* (bb) object detection results. Overall, TensorMask has a comparable box AP with Mask R-CNN and outperforms RetinaNet.

method	aug	epochs	AP^{bb}	$\text{AP}^{\text{bb}}_{50}$	$\text{AP}^{\text{bb}}_{75}$
RetinaNet, *ours*		24	37.1	55.0	39.9
RetinaNet, *ours*	✓	72	39.3	57.2	42.4
Faster R-CNN, *ours*	✓	72	40.6	61.4	44.2
Mask R-CNN, *ours*	✓	72	41.7	62.5	45.7
TensorMask, *box-only*	✓	72	40.8	60.4	43.9
TensorMask	✓	72	41.6	61.0	45.1

Table 4. **Object detection** *box* AP on COCO `test-dev`. All models use ResNet-50-FPN. 'TensorMask, *box-only*' is our model without the mask head: it resembles RetinaNet but with the mask-driven assignment rule and only 2 window sizes instead of 9 [23].

A.4. Mask-Only TensorMask

One intriguing property of TensorMask is that *masks are not dependent on boxes*. This not only opens up new model designs that are mask-specific, but also allows us to investigate whether *box predictions improve masks in a multi-task setting*. Here, we conduct experiments *without* the use of a box head. Note that although we predict masks densely, we still need to perform NMS for post-processing. If regressed boxes are absent, we simply use the bounding boxes of the masks as a substitute (and also to report box AP).

Tab. 5 gives the results. We observe a slight degradation switching from the default setting which uses original boxes (row 1) for NMS to using mask bounding boxes (row 2). After accounting for this, TensorMask *without a box head* (row 3) has nearly equal mask AP to the mask+box variant (row 2). These results indicate that the role of the box head is auxiliary in our system, in contrast to Mask R-CNN.

box head	NMS	AP	AP_{50}	AP_{75}	AP^{bb}	$\text{AP}^{\text{bb}}_{50}$	$\text{AP}^{\text{bb}}_{75}$
✓	bb	35.2	56.4	37.0	41.6	60.8	44.8
✓	mask-bb	34.9	56.0	36.7	39.7	59.1	41.8
	mask-bb	34.8	56.1	36.7	39.4	58.8	41.6

Table 5. **Multi-task benefits** of box training for mask prediction on COCO `val2017` with our final ResNet-50-FPN model.

A.5. Qualitative Comparisons and Calibration

We show more results in Figs. 10 and 11. For these, and all visualizations in the main text, we display all detections that have a *calibrated* score $\geq$0.6. We use a simple calibration that maps uncalibrated detector scores to precision values: for each model and for each category, we compute its precision-recall (PR) curve on `val2017`. As a PR curve is parameterized by score, we can map an uncalibrated score for the detector-category pair to its corresponding precision value. Score-to-precision calibration enables a fair visual comparison between methods using a fixed threshold.

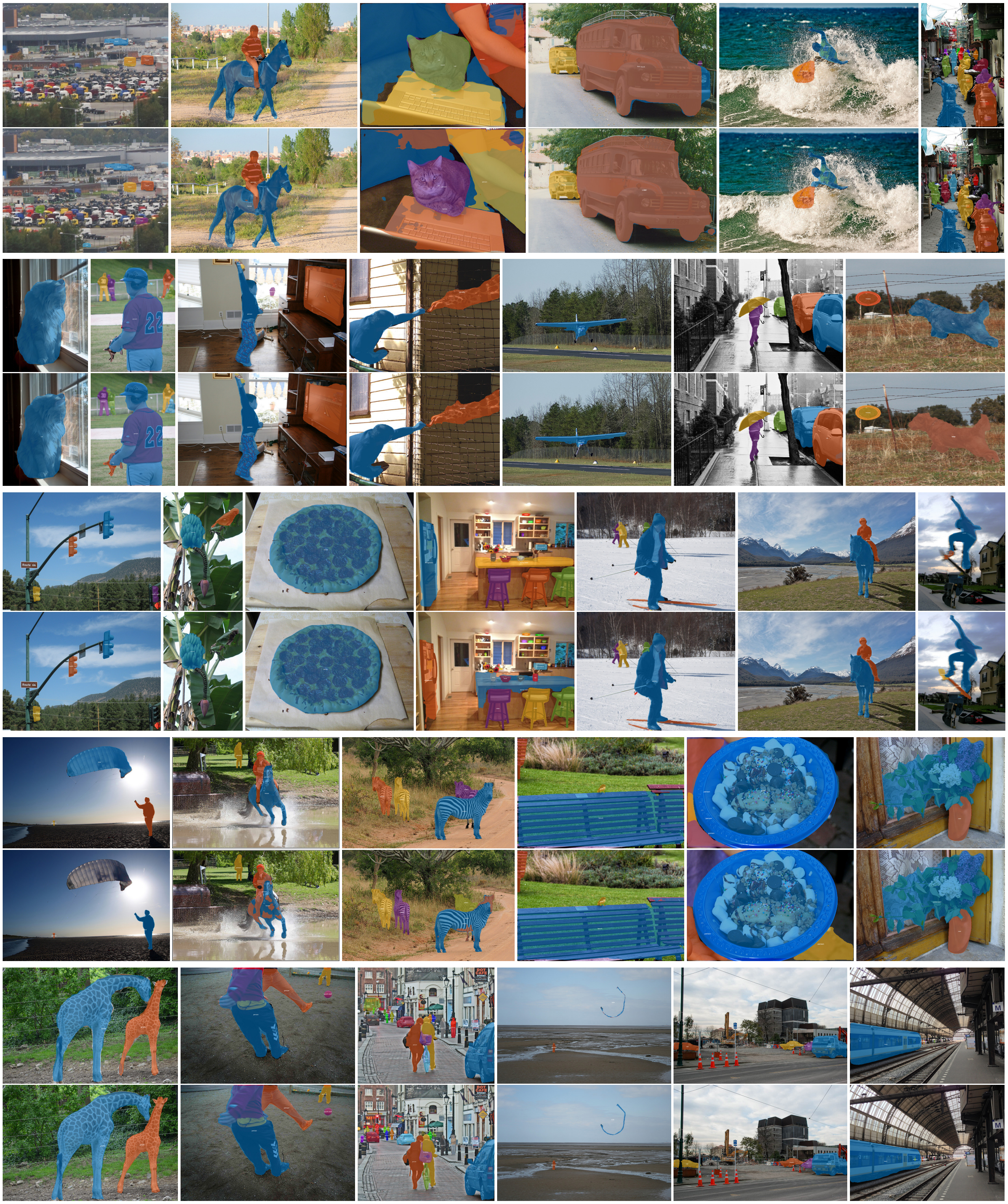

Figure 10. More results of Mask R-CNN [17] (top row per set) and TensorMask (bottom row per set) on the last 65 `val2017` images (continued in Fig. 11). These models use a ResNet-101-FPN backbone and obtain 38.3 and 37.1 AP, on `test-dev`, respectively. Visually, TensorMask gives sharper masks compared to Mask R-CNN although its AP is 1 point lower. Best viewed in a digital format with zoom.

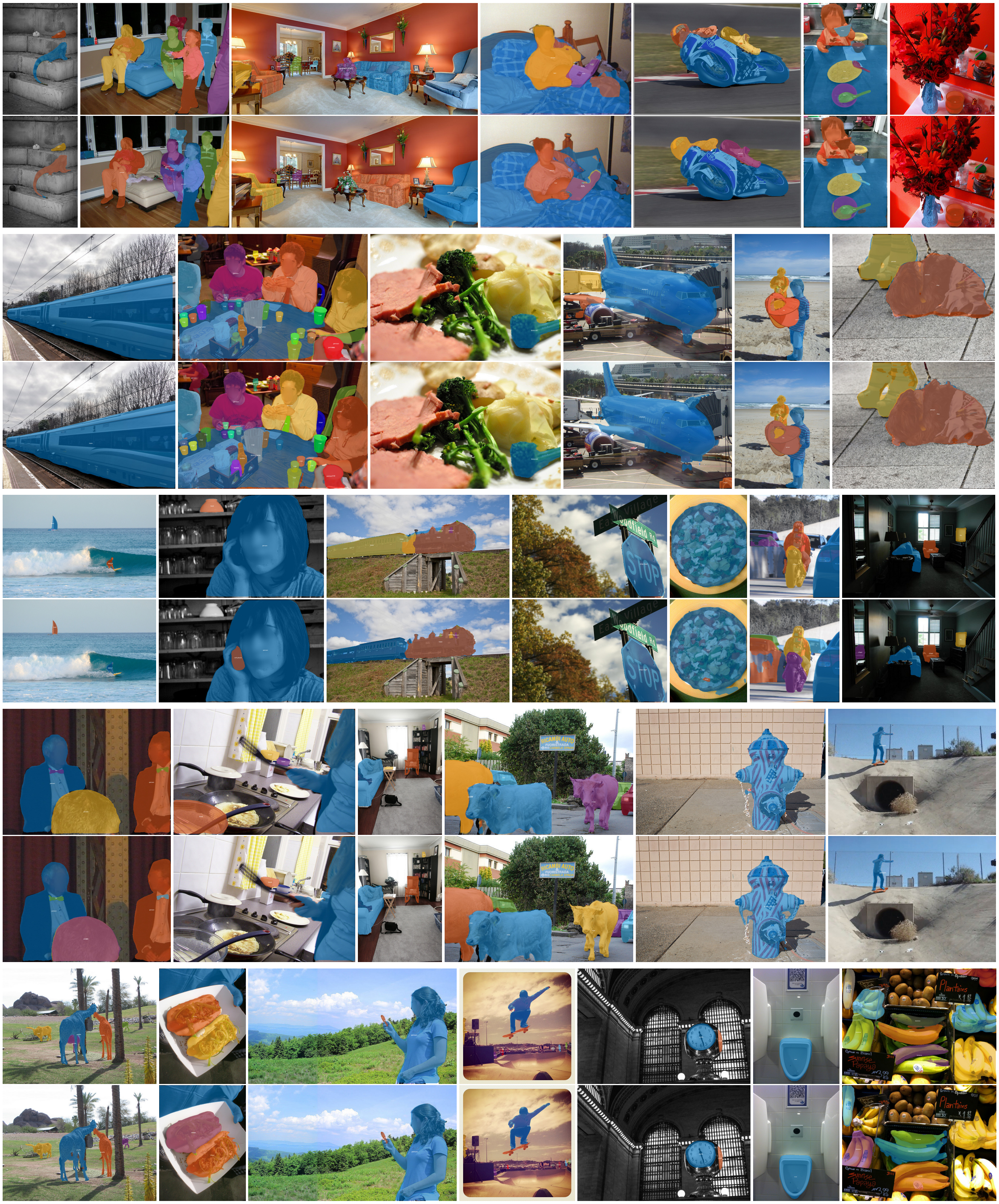

Figure 11. More results of Mask R-CNN [17] (top row per set) and TensorMask (bottom row per set) continued from Fig. 10.

References

[1] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *CVPR*, 2014. 2

[2] Anurag Arnab and Philip HS Torr. Pixelwise instance segmentation with a dynamically instantiated network. In *CVPR*, 2017. 3

[3] Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. In *CVPR*, 2017. 3

[4] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. *arXiv:1901.07518*, 2019. 2

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 3

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2, 3

[7] Jifeng Dai, Kaiming He, Yi Li, Shaoqing Ren, and Jian Sun. Instance-sensitive fully convolutional networks. In *ECCV*, 2016. 2, 3, 4, 7, 8, 9

[8] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016. 2

[9] Piotr Dollár, Zhuowen Tu, Pietro Perona, and Serge Belongie. Integral channel features. In *BMVC*, 2009. 1

[10] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010. 1, 4

[11] Ross Girshick. Fast R-CNN. In *ICCV*, 2015. 3

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2

[13] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018. 7, 8

[14] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv:1706.02677*, 2017. 7

[15] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *ECCV*, 2014. 2

[16] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. *arXiv:1811.08883*, 2018. 7, 8

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 1, 2, 4, 7, 8, 10, 11

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 7

[19] Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, and Carsten Rother. Instancecut: from edges to instances with multicut. In *CVPR*, 2017. 3

[20] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1989. 1, 2

[21] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *CVPR*, 2017. 2

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 4, 5, 6, 7

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 1, 2, 3, 5, 6, 7, 8, 9

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1, 2, 3, 7

[25] Shu Liu, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. SGN: Sequential grouping networks for instance segmentation. In *ICCV*, 2017. 3

[26] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018. 2

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *ECCV*, 2016. 1, 3, 5, 6, 7, 8

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 3

[29] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 2, 3

[30] Chao Peng, Tete Xiao, Zeming Li, Yuning Jiang, Xiangyu Zhang, Kai Jia, Gang Yu, and Jian Sun. MegDet: A large mini-batch object detector. In *CVPR*, 2018. 2

[31] Pedro Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *NIPS*, 2015. 2, 3, 6, 7

[32] Pedro Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *ECCV*, 2016. 3

[33] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. In *CVPR*, 2017. 3, 7, 8

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1, 2, 3, 6, 8

[35] Alexander Rush. Tensor considered harmful. 2019. 4

[36] R. Vaillant, C. Monrocq, and Y. LeCun. Original approach for the localisation of objects in images. *IEE Proc. on Vision, Image, and Signal Processing*, 1994. 1

[37] Koen EA van de Sande, Jasper RR Uijlings, Theo Gevers, and Arnold WM Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011. 2

[38] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001. 1

[39] Sergey Zagoruyko, Adam Lerer, Tsung-Yi Lin, Pedro Pinheiro, Sam Gross, Soumith Chintala, and Piotr Dollár. A multipath network for object detection. In *BMVC*, 2016. 2