You Better Look Twice: a new perspective for designing accurate detectors with reduced computations

Alexandra Dana
alex.dana@samsung.com
Maor Shutman
maor.shutman@samsung.com
Yotam Perlitz
ran.vitek@samsung.com
Ran Vitek
perlitz@gmail.com
Tomer Peleg
tomer.peleg@samsung.com
Roy J. Jevnisek
roy.jewnisek@samsung.com

Samsung Israel R&D Center
Tel Aviv, Israel

S1 Detection accuracy and computations of all published pedestrian detectors

Table S1 summarizes detection accuracy and computations measured by FLOPs of the latest published pedestrian detectors evaluated on the Citypersons reasonable validation dataset, on the original image size (2048×1024 px). Lower MR^{-2} indicates higher detection accuracy.

Table S1: Pedestrian detectors benchmark on Citypersons validation dataset. Detectors marked in bold are Pareto-efficient according to empirical findings in Figure 1 (main text). For models with published code FLOPs were directly measured, otherwise estimated using the Pedestron framework. The following backbone acronyms were used: MobileNet-V1(MV1); MobileNet-V2(MV2); ResNet-50(RN50); ResNet-50(RN101).

Detector	Measured?	GFLOPs	MR^{-2}
ALFNet (MV1)	✓	171	15.45%
RepLoss (RN50)	×	183	14.6%
Pedestron (MV2)	✓	262	10.24%
APD (DLA34)	✓	283	9%
PedHunter (RN50)	×	409	8.32%
Pedestron (HRNet)	✓	597	7.6%
CSP (MV1)	✓	352	15%
ALFNet (RN50)	✓	377	12.01%
CrowdHuman (RN50)	×	409	10.67%
CSP (RN50)	✓	657	11%
ACSP (RN101)	✓	731	7.63%
Citypersons (VGG16)	×	867	14.6%
AdaptiveNMS (VGG16)	×	867	11.9%
OR-CNN (VGG16)	×	867	12.8%
PSC-Net (VGG16)	×	867	10.5%
MGAN (VGG16)	×	894	11.3%
One-and-half (RN50)	N/A	N/A	8.12%

S2 BLT-net architecture

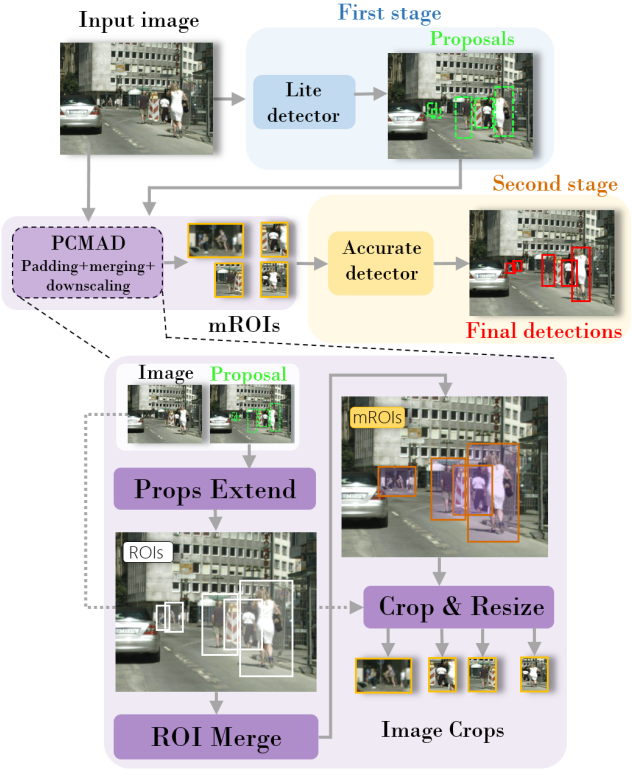

Figure S1: The input image is processed in the first-stage by a lite CNN with a high sensitivity to filter background regions. Resulting proposals are marked in dotted green rectangles. It is worth noticing that the scale and localization of detections at this stage are not perfect. Proposals are then padded (ROIs) and merged by the PCMAD algorithm. Based on the resulting mROIs, image crops are taken from the original image and further downscaled by the PCMAD algorithm, resulting the image crops with yellow margins. In this example the two left detected pedestrians were merged into a single mROI. The three right proposals are only donwscaled. Overall all image crops are downscaled to the same height and fed into the accurate second-stage detector for precise detections (marked with red rectanges). Figure best viewed in color.

S3 The CascadeMV2 net architecture

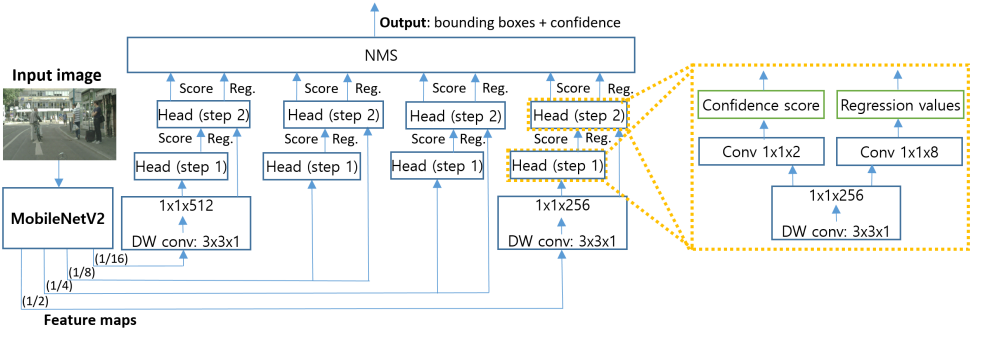

Figure S2: The CascadeMV2 architecture is based on concepts previously described [1, 2]. Feature maps at 1/2, 1/4, 1/8 and 1/8 of the input resolution are taken from the MobileNetV2 backbone. Some feature maps are further processed using depth-wise convolutions (DW; for reducing computations) to change their number of channels. Feature maps are then processed by a regression and confidence head, described in the yellow rectangle. First-step predictions are fine-tuned by the second-step head, that uses an identical design. The confidence score of the predicted bounding box and its values are taken solely from the second head. All predictions are then filtered using a NMS algorithm.

S4 The C&S net architecture

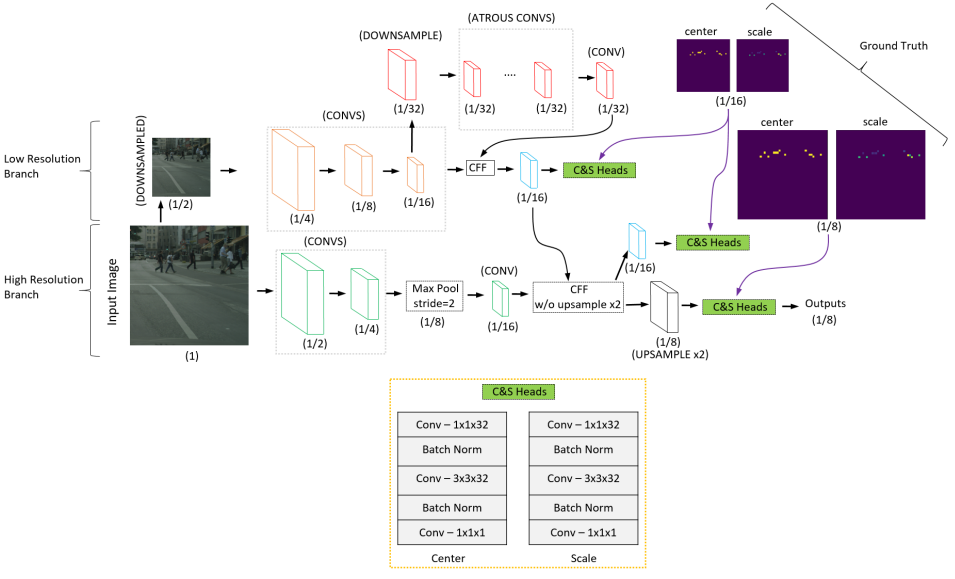

Figure S3: C&S net processes images using a shallow branch and a deep branch. The numbers in parentheses denote the output spatial sizes relatively to the original image size. The cascade feature fusion block (CFF) and additional implementation details regarding the blocks used in the shallow deep branches can be found in [3].

S5 First-stage resulting relative processed area

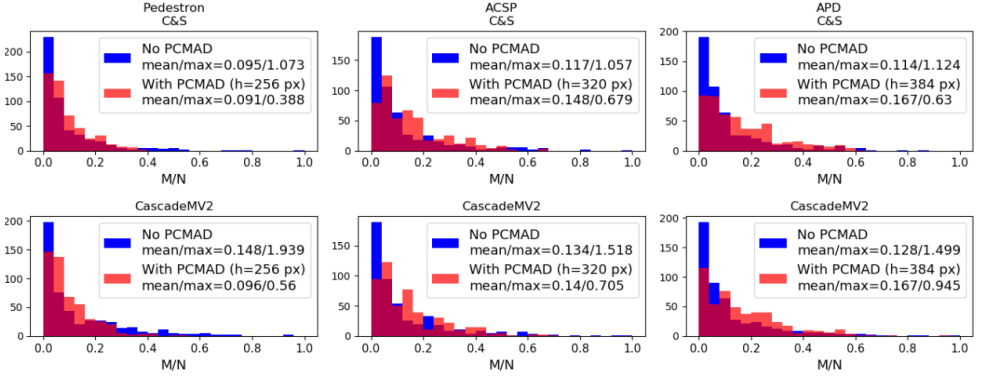

Figure S4: M/N (relative processed area) distribution per image for the Citypersons validation reasonable dataset, when using C&S net (top row) or CascadeMV2 net (bottom row) for the various evaluated pedestrian detectors. Red/blue colors depict the N/M distributions with/without applying the PCMAD algorithm. The *h* parameter indicates the specific heights image crops (based on the calculated mROIs) were downscaled to before applying on them the second stage detector.

S6 BLT-net additional misses

In this section we show all additional misses of BLT-net when using C&S net in the first-stage and Pedestron (HRNet) in the second-stage. Overall, BLT-net achieved a MR^{-2} lower by 1.79%. All additional misses resulting from the first-stage only are shown in Figure S5 and all additional misses resulting from the second-stage only are shown in Figure S6. These results show that misses from both stages are characterized by small and/or poorly illuminated or occluded pedestrians.

Figure S5: All first-stage additional misses for the Citypersons validation reasonable dataset, with respect to Pedestron (HRNet) applied on the entire image. The first-stage was implemented using the C&S net. Small bare and/or poorly illuminated pedestrians are labeled with (A) while partially occluded pedestrians are labeled with (B). All image crops have the same resolution and misses are marked with a red bounding box. The relative crop size with respect to a full image is shown in the top row.

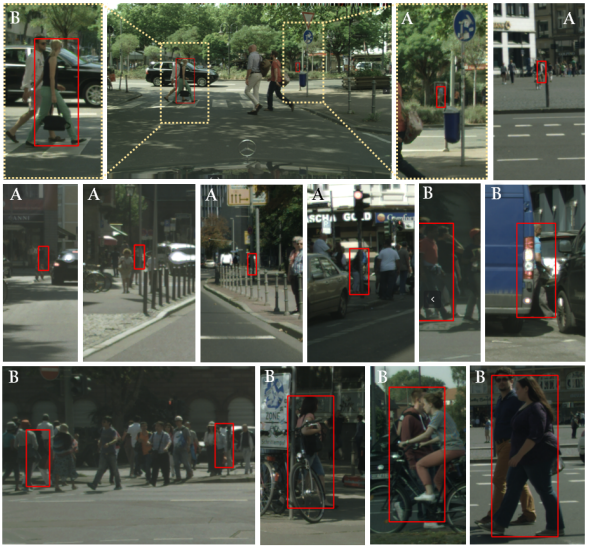

Figure S6: All second-stage additional misses for the Citypersons validation reasonable dataset, with respect to Pedestron (HRNet) applied on the entire image. The second-stage detector was implemented using the Pedestron (HRNet) detector. Small bare and/or poorly illuminated pedestrians are labeled with (A) while partially occluded pedestrians are labeled with (B). All image crops have the same resolution and misses are marked with a red bounding box. The relative crop size with respect to a full image is shown in the top row.

References

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

[2] Wei Liu, Shengcai Liao, Weidong Hu, Xuezhi Liang, and Xiao Chen. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 618–634, 2018.

[3] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–420, 2018.