Occlusion-aware pedestrian detection

Apostolopoulos, Christos; Nasrollahi, Kamal; Yang, Ming-Hsuan Yang; Jahromi, Mohammad Naser Sabet; Moeslund, Thomas B.

Published in:
Eleventh International Conference on Machine Vision, ICMV 2018

DOI (link to publication from Publisher):
10.1117/12.2523107

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Occlusion-aware pedestrian detection

Apostolopoulos, Christos, Nasrollahi, Kamal, Yang, M. Hsuan , Jahromi, Mohammad N. S., Moeslund, Thomas

Christos Apostolopoulos, Kamal Nasrollahi, M. Hsuan Yang, Mohammad N. S. Jahromi, Thomas B. Moeslund, "Occlusion-aware pedestrian detection," Proc. SPIE 11041, Eleventh International Conference on Machine Vision (ICMV 2018), 110410I (15 March 2019); doi: 10.1117/12.2523107

SPIE.

Event: Eleventh International Conference on Machine Vision (ICMV 2018), 2018, Munich, Germany

Occlusion-aware Pedestrian Detection

Christos Apostolopoulos[1], Kamal Nasrollahi[1], M. -Hsuan Yang [2], Mohammad N. S. Jahromi[1], Thomas B. Moeslund[1]
[1]Visual Analysis of People Laboratory, Aalborg University, Denmark
[2] Department of EE-CS, University of California at Merced, USA

ABSTRACT

Failure in pedestrian detection systems can be extremely crucial, specifically in driverless driving. In this paper, failures in pedestrian detectors are refined by re-evaluating the results of state of the art pedestrian detection systems, via a fully convolutional neural network. The network is trained on a number of datasets which include a custom designed occluded pedestrian dataset to address the problem of occlusion. Results show that when applying the proposed network, detectors can not only maintain their state of the art performance, but they even decrease average false positives rate per image, especially in the case where pedestrians are occluded.

1. INTRODUCTION

Object detection is a fundamental and important area of computer vision with many applications such as surveillance systems, automatic driven vehicles and human-computer interaction, where an object of interest is attempted to be located or detected over a series of images. Common challenges in this area involve being able to robustly detect an object under varying situations such as camera distances, angles, object perturbations and weather conditions. One of the most common detection objects of interest include pedestrians, where in a street with a variety of challenging situations one or more pedestrians are attempted to be detected. Several approaches have been proposed over the last decade [1]–[8], which addresses some of those challenges and achieve certain performance.

Even though performance has significantly increased, each method still suffers and performs poorly in certain situations. Generally, when dealing with pedestrians, there is a number of challenging factors that can decrease precision. These include situations such as occlusion (when the pedestrian is occluded by another pedestrian or an external object), background clutter (when the background has similar texture to the pedestrian) or low resolution (when the pedestrian is of a small size - usually under 30-pixel height - [9]). When a detector fails to enclose a pedestrian (such as in Fig. 1) correctly, there is no way to recover the detector result as it just output the incorrect result. Implementing a system which is able to recognize these failures effectively and recover (if possible) the detectors estimation would increase accuracy and precision by a large margin.

Due to the rising interest in convolutional neural networks (CNN) and their constant increase in performance (and out- performance in relation to state-of-the-art) in fields such as classification, segmentation and recognition [10], in this paper, we propose to re-evaluate pedestrian detector output to detect its possible failures, refine its result and hence increase overall performance through an end-to-end fully convolutional neural network (FCNN) for pixel-wise pedestrian prediction. FCCNs do not rely on fully connected layers (as opposed to CNN) and all the learnable layers are convolutional. This makes FCNN an effective structure to apply any input size or image resolution where image local features can effectively extract spatial information. We attempt to focus on the problem of occlusion. Specifically, since occlusion has been shown to heavily affect detector performance [11], we construct our own custom pedestrian segmented occluded dataset. We show that by training such a network, we can estimate failures based on the challenging factors mentioned above or from drawbacks that each detector may have.

The ***main contributions*** of this work are the followings:

- We construct a novel end-to-end fully convolutional neural network for pixel-wise pedestrian prediction using decon- volutional layers and concatenating earlier and later layers.
- We construct the first partial occluded segmented pedestrian dataset for learning pedestrian behavior.
- We propose a novel approach for re-evaluating detector results and increasing performance, especially in the problem of occlusion.

Eleventh International Conference on Machine Vision (ICMV 2018), edited by Antanas Verikas, Dmitry P. Nikolaev, Petia Radeva, Jianhong Zhou, Proc. of SPIE Vol. 11041, 110410I

In the next sections, we briefly discuss about related work in the literature.

2. RELATED WORK

- **Classical method**: Majority of previous works concerning pedestrian detection are based on image segmentation [12]. Image segmentation that relies on graph theory such as minimum spanning tree (MST) or active shape model based, such as ped-cut are example of this group [13], [14]. Furthermore, traditional pedestrian detectors, Viola and Jones paradigm such as such as ACF [15] or Integral Channels Features (ICF) [16] are widely used in this context.

Figure 1. An example detector failure for the pedestrian of interest. Green box denotes detector result where red box denotes ground truth.

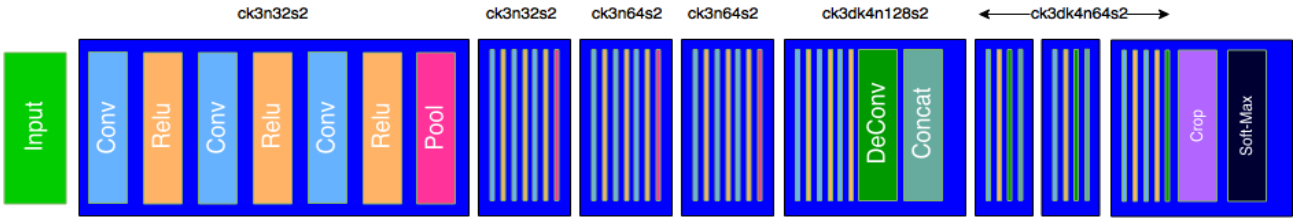

Figure 2: The proposed network architecture. ck denotes convolutional kernel, dk deconvolutional kernel, n number of output (feature maps) and s stride.

- **CNN-based method**: Inspired by deep convolutional neural network, CNN-based models [17], [18], YOLO [19] and fast R-CNN method [20] have pushed pedestrian detection results to an unprecedented level.
- **Fully CNN-based method**: Samples include the works in [21], [22]. Since in this work, the pedestrian prediction is done on a pixel-wise level (spatial information), the proposed design is therefore associated with this class which is more suitable method in this scenario. We discuss the proposed network in coming section.

3. PROPOSED FCNN NETWORK

In this section we describe the design procedure of the proposed system which is twofold between the training and testing of the system.

3.1. Training

- **Architecture**: The proposed network is comprised of eight convolutional layers, four de-convolutional layers, one crop layer and a softmax loss layer (Fig. 2). The first five convolutional layers are convolved with a set of (3×3) filters size, and follow a very typical convolutional neural network pattern similar to [17], where after each convolution a rectified linear unit (ReLu) activation function and pooling is applied. Note that the fifth convolutional layer does not have a max pooling layer due to its input dimensions being already small in spatial size (12×12). After the up-sampling takes place, each pair of deconvolutional and convolutional

layers that have the same spatial size are concatenated as illustrated in Fig. 3. The reasoning behind this is when dealing with low resolution images and segmentation methods, it is important to combine both early (corners, edges) and late (semantic) layer information which is extracted by the network. Furthermore, the concatenation deals with information that can be lost at the images boundaries due to deep convolutions (small feature maps). Following each de-convolutional layer a convolutional layer is placed that uses a filter size of one as a dimensionality reduction tool. The dimensionality is finally reduced in the final deconv layer to 2 in order to match our label classes. Due to the up-sampling in the last layer slightly exceeding the label size, a cropping layer is added. Note that the cropping is small in size (2x2) and thus should not affect performance. These are then fed into the softmax loss function in order to compute the networks loss.

The network parameters during training include a *learning rate* of 10^{-9} over $100k$ total iterations. The *average loss* is reported every 20 iterations and *test interval* of 500 iterations for validation. The weight initialization for the convolution layer is done by using a Gaussian distribution while a xavier initialization [23] is used for the de-convolution layer. The proposed network is also pre-trained on the icome dataset [24] for cloth parsing by mean subtraction of its train/ test images to increase the network accuracy.

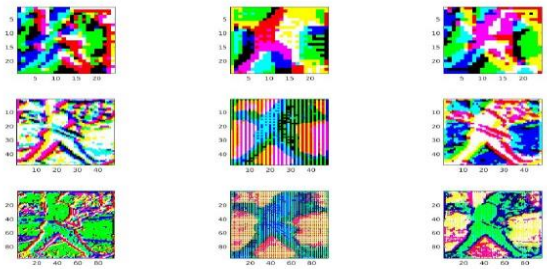

Figure 3. Up-sampling/concatenation procedure. From top to bottom: fifth, sixth and seventh layer. From left to right: Convolutional layer, de-convolutional layer and concatenated layer.

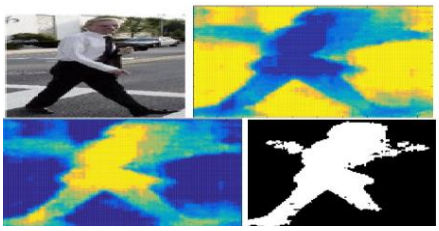

Figure 4. Top left:input image. Top Right:Background Heatmap. Lower left: Foreground Heatmap. Lower right: Fused Result.

- **Re-size/Distort data**: All training data is labeled in a pixel-wise binary classification manner (Background vs Fore- ground). The size of the input image to the network is another significant matter. Since pedestrian images are mostly captured with surveillance cameras, there should be a compromise of the input size in order to take into account low and high resolution images. The input size in this work has been chose to be (192 × 192). The datasets used contain images in a good resolution since it is important to capture details during training. The resolution cannot be expected to be of the same quality in test time though, therefore both the training and testing data is resized in two ways: First a bicubic interpolation takes place and each image is resized to a (40 × 80) and then afterwards a nearest neighbor algorithm is used in order to upsample the bicubic resized image to the desired (192 × 192) size respectively. In this way, both high and low resolution images will be re-adjusted during train/test phase.
- **Augment data**: To avoid over-fitting the network while keeping a large amount of variant data, we augment our data in terms of translation (30% left, 30% right, 30% upwards and 30% left and upwards combined), rotation (between −7 and 7), scaling (ranges from 1.6 to 1 with step-down size of 0.18) and flipping.

3.2. Testing

- **Extract bounding box coordinates**: The network relies on a pedestrian detector in order to receive an initial estimation, which is given in terms of image coordinates and a bounding box size.
- **Scaling**: The assumption is made that the detector may not perform well in certain situations. Hence a scaling factor is applied to every bounding box in order to assure that if the measurement is incorrect, there will be a margin that will assure that the pedestrian is included in the bounding box area.
- **Pre-process inputs/Pass through network**: Pre-processing is done through scaling and mean subtraction techniques as similar as in offline mode. The network returns two (192 × 192) heatmaps for the foreground and background estimation. These are then fused in order to return the final result, which is a binary image where 1 represents a pedestrian and 0 the background. The pixel decision is made by looking at which channel has the highest value for the given pixel.This can be seen in Fig. 4.
- **Calculate new bounding box coordinates/Re-adjust result**: After the binary image is constructed, the final refined estimation is made. By extracting the biggest BLOB (to minimize estimation noise), the coordinates of the bounding box are measured with respect to three spaces: the (192 × 192) network estimation, the original bounding box (prior to resize) and the global input image. The new result can then be plotted on the image and compared with the ground truth and detector estimation.

4. DATABASE

- **Pre-Training**: As discussed earlier, the icome dataset [24] is used to pre-train the network to learn a how an overall human shape should look like.
- **Training**: A collection of datasets are collected before-hand in order to train the network efficiently. These include positive examples (datasets of pedestrians images and their segmented labels) and negative examples (datasets of non pedestrian images and their segmented labels). This collection consists of three dataset. Penn-Fudan [25] and Daimler Pedestrian Segmentation [14] are two well-known pedestrian datasets that are used to form our collection. The third dataset, in particular, is considered for partially occluded segmented pedestrian scenario. To best of our knowledge, there does not exist publicly available a partial occluded pedestrian segmentation dataset. In this work, the dataset in [26] in order to deal with occlusion (≈ 500 partially occluded subjects).
- **Test**: We tested our proposed system on four widely used datasets for pedestrian detection problem. But, in this work, we tend to report the results obtained on challenging caltech dataset [27].

5. SIMULATION RESULTS

To evaluate the proposed network, we apply six state-of-the-art pedestrian detectors on the Caltech dataset and use their results as an input into the network. Specifically, we use MultiFtr+CSS - MultiFtr+Motion [28], ChnFtrs [16], FPDW [29], ACF+Inria [15] and FeatSynth [30]. The reason for using this specific dataset is due to that it is the only dataset containing isolated sub-sets of its dataset that empasize on scale and occlusion variants, which allow us to further investigate the proposed networks result in specific situations.

As a metric for evaluation before and after applying the network, we adopt the test setup in [9], where log-average miss rate versus false positives per image is plotted by thresholding the confidence score.

5.1. Data Pre-processing and Filtering

For each image, every detector returns a number of results, where multiple results can correspond to the same pedestrian. For this reason we filter detections and attempt to link one detection per ground truth image. This is done by calculating overlap precision, and a detection is matched if its precision is over 0.5. Detections with the highest condence score are matched first since they are most likely to be good measurements.

The assumption is made that the detector may not perform well in certain situations, therefore a scaling factor is applied to every bounding box returned from the detectors results to maximize the features of the pedestrian included

in the region of estimation. Note that these bounding boxes are cropped from the benchmarks images after the scaling takes place and the initial detection is only used as an approximation. The scaling factor found to maximize our networks output score was calculated to be 1.2, which can be contributed to the fact that since we choose detection with overlaps already significantly high ($\geq$ 0.5), then we do not need to scale the bounding box too much in order to cover the region that visualizes the pedestrian while at the same time minimizing the amount of noise being introduced.

5.2. Network Confidence Score

The proposed network does not return a confidence score but only an approximation on where the pedestrian is located, and since most pedestrian evaluation metrics takes this score into account, we use the Aggregated Channel Features detectors confidence score trained on the Caltech data set for evaluation. The reason behind using the Caltech trained detector for this task is due to the fact that the sliding window size that the authors chose when training is smaller than the INRIA one (64x32 versus 128x64). It is also worth noting that the detector is trained on full-scene images and not cropped bounding boxes. Having in mind that this confidence does not describe the networks approximation, we calculate the percentage that the detectors bounding box covers in the networks output and the resulting score is used as the final confidence.

5.3. Results

We use a combination of pedectors that were build specifically for pedestrian detection (MultiFtr+CSS, MultiFtr+Motion, ChnFrts, FPDW) and other object detection (FeatSynth, ACF+Inria). We notice that the overall best performance across all datasets comes from the most recent work relating pedestrian detectors (MultiFtr+CSS and MultiFtr+Motion).

As can be seen in Fig .5, The proposed network can achieve state-of-the-art performance over reasonable samples and decreases false positives per image about 4.5% on average through all algorithms, where it works most effectively on the MultiFtr+CSS algorithm (8% decrease). We notice that the worse that the performance of the detector is, the smaller the contribution of the proposed network (2% at ACF+Inria which has 84% false positives and 8% at MultiFtr+CSS which has 61%). This can be tracked back to the requirement we have that the network should receive a reliable measurement as an input.

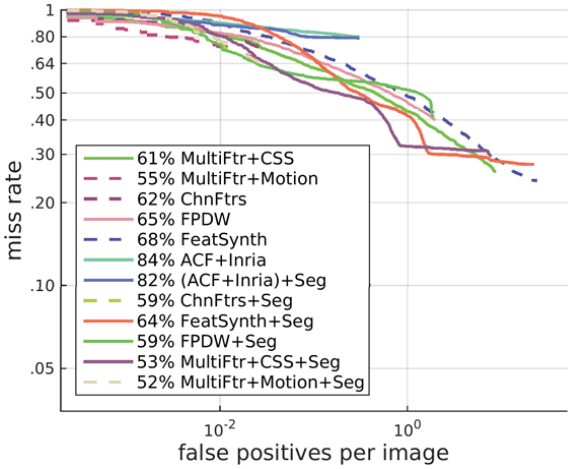

Figure 5. Measurement on Caltech dataset (reasonable samples).

Scale is divided into three categories : near (80 or more pixels), medium (30-80 pixels) and far (30 pixels or less). From Fig .6, the proposed network can achieve state-of-the-art performance over scale variant samples and decreases false positives per image about 1.5% on average through all algorithms on near scale, 3.8% on medium scale and -1.1% on far scale. All detectors performances suffer heavily on far scale pedestrians (nearly 100% false positives per image), so contribution on our side is nearly impossible and not expected. As it can be observed also in the reasonable samples, the false positive samples percentage that we can expect the biggest contribution is around the range of 40-80% , which in this example is provided both in the near and medium scale on different detectors (e.g. 8% decrease on FeatSynth in near scale and 7% on MultiFtr+Motion in medium scale).

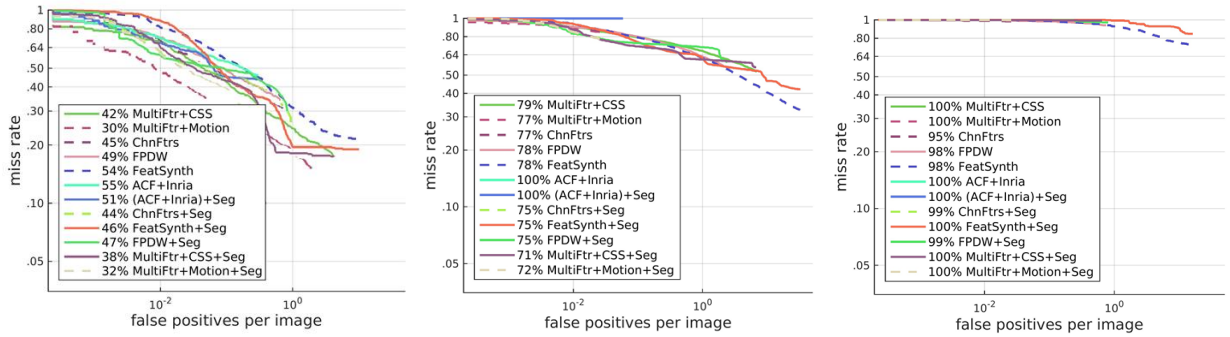

Figure 6. Measurement on Caltech dataset (scale variant samples). From left to right - near, medium and far scale.

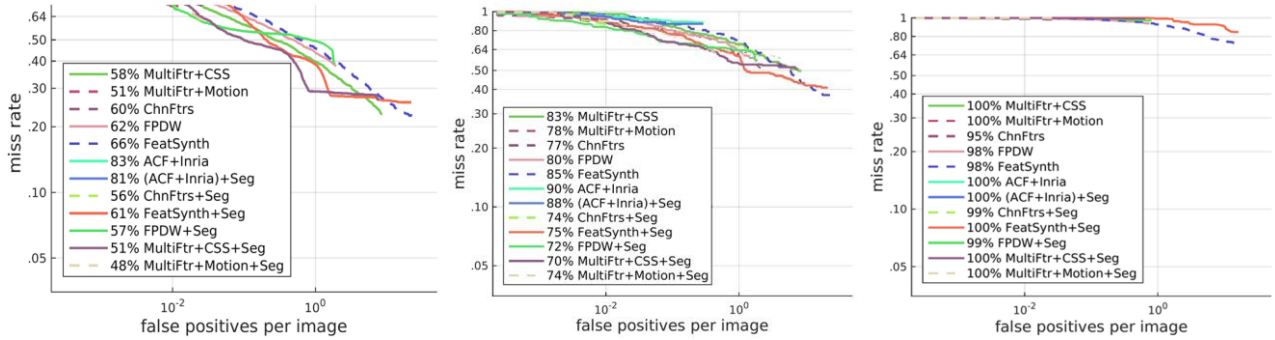

Figure 7. Measurement on Caltech dataset (occluded variant samples). From left to right - no, partial and heavy occlusion.

Occlusion is divided into No Occlusion (0% occluded area), Partial occlusion (1-35% occluded area) and heavy occlusion (36-80% occluded area). From Fig. 7, the proposed network can achieve state-of-the-art performance over occluded variant samples and decreases false positives per image about 4.3% on average through all algorithms on no occlusion, 6.6% on partial occlusion and 2.3% on heavy occlusion. The custom dataset made on partial occlusion's impact can be seen on these results. We observe a bigger percentage drop on this plot than in the other sets in the higher ranges (70-85% false positives per image). On the other hand in heavy occlusion, even though information is very limited, there is still a minor increase, which can prove that our custom dataset can show promising results in understanding where the representation of one pedestrian ends, even if that is limited.

6. CONCLUSION

In this paper, we propose a fully convolutional neural network for refining results for pedestrian detection (using the output of detectors), while being able to handle occluded pedestrians. We configure the design of our network by concatenating earlier and later de-convolutional and convolutional layers in order to precisely segment where a pedestrian is located. For dealing with partial occlusion, we construct our own custom pedestrian dataset to further learn our network about variations in pedestrian appearance. Results show that when applying the proposed network, detectors can not only maintain their state-of-the-art performance but even decrease average false positives per image, especially in the case where pedestrians are occluded.

REFERENCES

[1] Q. Ye, T. Zhang, Q. Qiu, B. Zhang, J. Chen, and G. Sapiro, "Self-learning scene-specific pedestrian detectors using a progressive latent model," *CoRR, abs/1611.07544*, vol. 2, 2016.

[2] S. Huang and D. Ramanan, "Expecting the unexpected: Training detectors for unusual pedestrians with adversarial imposters," in *IEEE, CVPR*, vol. 1, 2017.

[3] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," *arXiv preprint arXiv:1704.02431*, 2017.

[4] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multiperson pose estimation in the wild," *arXiv preprint arXiv:1701.01779*, vol. 8, 2017.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[6] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 32–39.

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[8] Z. Lin and L. S. Davis, "A pose-invariant descriptor for human detection and segmentation," in *European Conference on Computer Vision*. Springer, 2008, pp. 423–436.

[9] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[10] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognition*, 2017.

[11] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *Computer Vision, 2005. ICCV 2005*, vol. 1. IEEE, 2005, pp. 90–97.

[12] C. T. Zahn, "Graph-theoretical methods for detecting and describing gestalt clusters," *IEEE Transactions on computers*, vol. 100, no. 1, pp. 68–86, 1971.

[13] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International journal of computer vision*, vol. 59, no. 2, pp. 167–181, 2004.

[14] F. Flohr, D. Gavrila *et al.*, "Pedcut: an iterative framework for pedestrian segmentation combining shape models and multiple data cues." in *BMVC*, 2013.

[15] P. Dolla ŕ, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[16] P. Dolla ŕ, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," 2009.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[18] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *ECCV*. Springer, 2016, pp. 354–370.

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016, pp. 779–788.

[20] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast r-cnn for pedestrian detection," *IEEE Transactions on Multimedia*, 2017.

[21] O. Matan, C. J. Burges, Y. LeCun, and J. S. Denker, "Multi-digit recognition using a space displacement neural network," in *Advances in neural information processing systems*, 1992, pp. 488–495.

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[23] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[24] Z. Wu, Y. Huang, Y. Yu, L. Wang, and T. Tan, "Early hierarchical contexts learned by convolutional networks for image segmentation," in *Pattern Recognition (ICPR)*. IEEE, 2014, pp. 1538–1543.

[25] L. W. et al, "Penn-fudan database for pedestrian detection and segmentation."

[26] J. Mar´ın, D. Va´zquez, A. M. Lo´pez, J. Amores, and L. I. Kuncheva, "Occlusion handling via random subspace classifiers for human detection,"*IEEE transactions on cybernetics*, vol. 44, no. 3, pp. 342–354, 2014.

[27] P. Dolla´r, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 304–311.

[28] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *CVPR, 2010 IEEE conference on*. IEEE, 2010, pp. 1030–1037.

[29] P. Dolla´r, S. J. Belongie, and P. Perona, "The fastest pedestrian detector in the west." in *Bmvc*, vol. 2, no. 3. Citeseer, 2010, p. 7.

[30] A. Bar-Hillel, D. Levi, E. Krupka, and C. Goldberg, "Part-based feature synthesis for human detection," in *ECCV*. Springer, 2010, pp. 127–142.