Intl. Conference on Pattern Recognition, 2016

Detection and Localization with Multi-scale Models

Eshed Ohn-Bar and Mohan M. Trivedi
Computer Vision and Robotics Research Laboratory
University of California San Diego
{eohnbar, mtrivedi}@ucsd.edu

***Abstract*—Object detection and localization in images involve a multi-scale reasoning process. First, responses of object detectors are known to vary with image scale. Second, contextual relationships on a part-level, object-level, and scene-level appear at different scales of the image. This paper studies efficient modeling of these two components by training multi-scale template models. The input to the proposed algorithm involves image features computed at varying image scales, hence operating on volumes in the feature pyramid. The approach generalizes single-scale, local-region detection approaches (e.g. sliding window or region proposals), jointly learning detection and localization cues. Extending the common single-scale detection to a multi-scale volume allows learning scale-specific models as well as analyzing the importance of contextual information at different scales. Experimental analysis on the PASCAL VOC dataset shows the method to considerably improve both detection and localization performance for different type of features, histogram of oriented gradients and deep convolutional neural network features.**

I. Introduction

Modeling information at different image scales is fundamental to many tasks in computer vision [1]–[3]. In object detection, detection at different scales is commonly achieved using models that are trained for classification of cues inside a local region (e.g. sliding window or region proposals). This popular scheme is often applied over re-sampled versions of the original image in order to handle detection at multiple scales, which implies evaluation of the trained model at different scales independently. A key disadvantage of such an approach is that it ignores the highly patterned cues that manifest at different scales and resolutions. Looking at Fig. 1, the information over scales is far from independent. This work proposes a novel formulation of the multi-scale detection pipeline, generalizing the traditional single-scale approach to reason over detection and localization cues found in all of the scales of the sampled image pyramid. Our main contribution is in a learning framework that can better leverage patterns and structure in multi-scale cues, hence we refer to the proposed approach as **MSS** (Multi-Scale Structure).

Fig. 2 depicts the proposed MSS detection pipeline. The learned weights for one of the MSS templates (for a car model) are shown as well, with a corresponding example positive image sample. The model shown in Fig. 2 was trained using conv_5 convolutional feature maps extracted using AlexNet [4]. As shown, cues are often selected at high amount in the scale of best match (second scale from the left in the figure), but the overall cue selection process spans all of the scales of a 7-scale image pyramid used in the experiments. This is intuitive - although only one of the scales best fits the car, different scales may contain cues important for the localization and detection of that vehicle in the best fitting scale, such as parts of the vehicle (the bumper, license plate, or tail lights) and contextual scene information (such as road cues, or other objects). Modeling such multi-scale information is useful for detection and localization, leading to significant gains in detection performance on the challenging PASCAL VOC object dataset [5].

Fig. 1: Traditional object detectors train and test models at a single scale, thereby ignoring information over scales at a specific spatial window location. Our study is motivated by the fact that multi-scale information is highly structured. For instance, cues at different scales, such as the road or the license plate, provide detection and localization information for the car. The overlap with the ground truth box (shown in red) also follows a clear structural pattern.

This study presents significant gains in detection performance can be obtained without altering the underlying descriptor but by replacing the traditional multi-scale pipeline with the proposed novel multi-scale structure MSS approach.

II. The Multi-scale Structure (MSS) Approach

Multi-scale reasoning extracted from multiple image scales has been shown to be essential for a variety of vision tasks (e.g. image segmentation [6]), yet its usage in object detection has been limited. Two main differences stand between the proposed, MSS approach, and related studies employing multi-scale contextual reasoning. First, MSS classifies all scales in a feature pyramid at once, while related studies classify them independently [7]–[19], often with models learned over a single image scale. For instance, the deformable part model [9] employs part HOG [20] features extracted from twice the resolution scale of the root template model, yet the main multi-scale sliding window pipeline is left unchanged, resulting in limited multi-scale reasoning capabilities. Other approaches, such as R-CNN [7], SPPnet [8], or DeepPyramid [21], also operate on individual local region within a single image scale, as opposed to features in regions across multiple scales of an image/feature pyramid. Second, the MSS framework proposes a modification to the inference label space, so that both a detection label and a scale label are predicted. This procedure

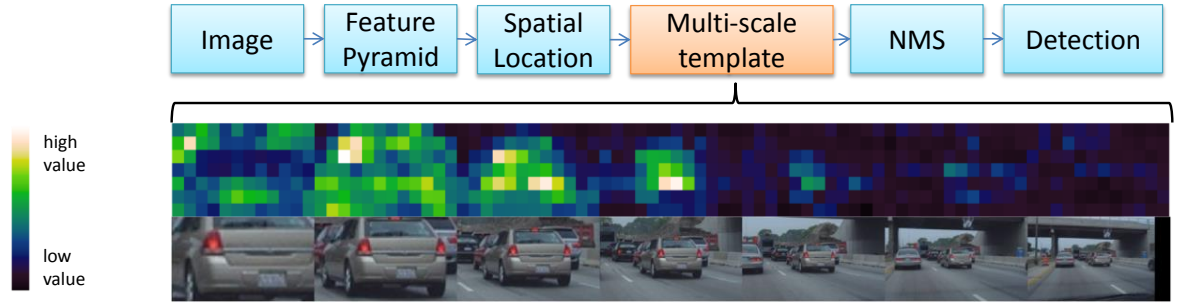


Fig. 2: The proposed multi-scale structure (**MSS**) approach learns joint detection and localization multi-scale templates by operating on volumes in a feature pyramid. An example MSS model with CNN features (one out of the 7 trained for the car object category) is visualized with a corresponding positive image sample. For a scale-specific car detection task, cues at different image scales, both adjacent and remote, are shown to be useful (in the visualization, brighter colors imply spatial locations with greater discriminative value).

leverages localization cues across the scales, and thereby differs from the studies of [6], [22], [23], where there is no such modification to the label space.

In this section, we outline the mathematical formulation of the proposed MSS approach. We demonstrate traditional object detectors to be a special case of our generalized all-scale framework.

A. *Feature pyramid*

We pursue two efficient approaches for obtaining a feature pyramid. Two types of commonly employed features are employed in order to study generalization of the proposed MSS approach. First, we employ the HOG implementation of [9] which is still widely used and serves as a classical baseline. Because HOG features are sensitive to scale, we demonstrate large detection performance gains by employing the proposed MSS approach over the single-scale baseline. We also utilize richer deep Convolutional Neural Network (CNN) features [7], [21]. Motivated by existing CNN-based object detectors, [11], [24], an efficient pyramid is extracted with a truncated version of the 8-layer AlexNet [4] network which won the ILSVRC-2012 ImageNet challenge. We employ the fifth convolution layer, which outputs 256 feature channels. The input to each convolutional or max pooling layer is zero-padded so that the features in a zero-based pixel location (x, y) in the feature space were generated by a receptive field centered at $(16x, 16y)$ in the image space (a stride of 16). As noted by [21], the CNN features already provide part and scale selective cues. The conv_5 features are enhanced by employing a 3×3 max-poolying layer with stride of 1. For direct comparison with [21], the same feature extraction and pyramid pipeline was implemented.

B. *Single-scale models for object detection*

The detection baseline employs a fixed size model and a pyramid of features in order to handle detection at multiple scales (see Fig. 3). Let $p_s = (x, y, s)$ be a window in the s-th level of a feature pyramid with S scales anchored in the x, y position. For now, we assume a single aspect ratio model for simplicity but training different aspect ratio models will be studied in the experimental analysis. Generally, the feature image is at a lower spatial resolution than that of the original image. Consequently, a zero-based index (x, y) in the feature map can be mapped to a pixel in the original image using a scale factor (cx, cy) based on the resolution of the feature map (for HOG, $c = 8$, and for CNN-conv_5 features, $c = 16$). Mapping spatial locations across scales can be achieved by a multiplication by the scale factor.

Given a local window of features, $\phi(p_s) \in \mathbb{R}^d$, the model learns a classification scoring function (in our case, a Support Vector Machine - SVM [25])

$$f(p_s) = w \cdot \phi(p_s) \tag{1}$$

The model size is an overhead parameter, fixed according to the size of the smallest object to be detected. Under this formulation training a model involves only the features in the local window, which is quite limited. As a matter of fact, even humans may have trouble identifying objects from background from cropped local windows. Because both training and testing involve classification of local windows in a single-scale ($\phi(p_s)$), testing must involve repeated classification of the same spatial location in the image pyramid across different scales. Finally, as commonly performed in state-of-the-art models, the scored windows are resolved using a heuristic non-maximum suppression module, which does not reason over image feature responses, multi-scale information, object and scene relationships, and more.

An improvement over this approach has been described using a template pyramid approach, which can be described with nearly identical notation. Several studies employ template pyramids [13], [16], [17], [26], [27] as it was shown to improve detection performance due to capturing scale-specific cues. For instance, larger objects contain more detailed information which can be leveraged for improved classification. Here, a differently-sized detection template is trained for each scale in

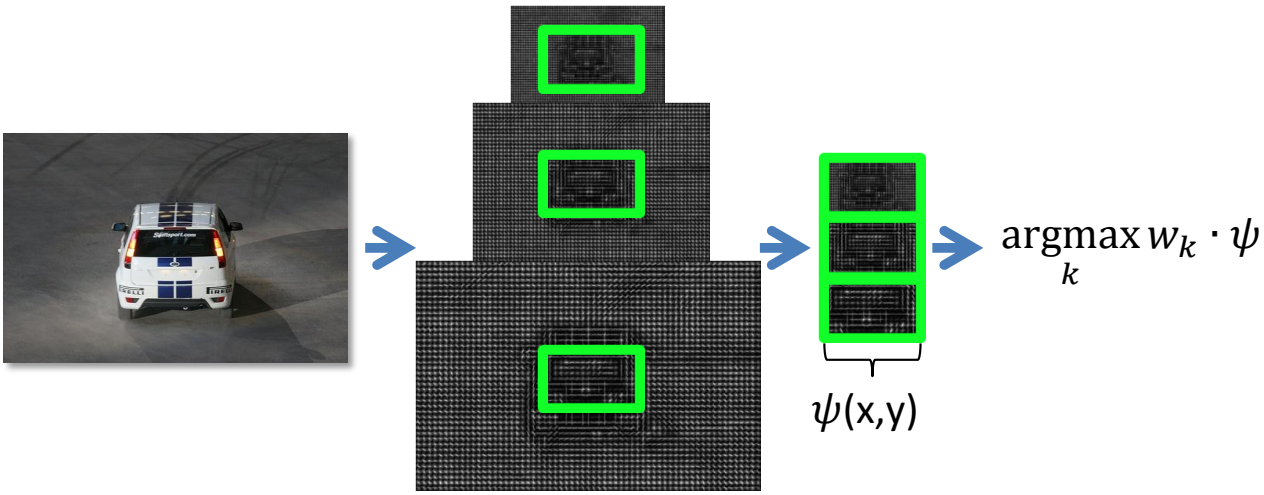


Fig. 3: Each MSS template w_k operates on the scale volume ψ and predicts a score of confidence for the presence of the object in a scale class, k. In inference, the highest score among the MSS templates is used for placing the final detection box in a certain scale.

the image pyramid, $(w_1, \ldots, w_S)$. In inference with a template pyramid, each spatial location p in the image (note that no re-scaling of the image is needed so that the s subscript is dropped) is classified with multiple scale-specific templates,

$$f(p) = \max_{s \in \{1,\ldots,S\}} w_s \cdot \phi(p) \tag{2}$$

Although modeling scale-specific cues has been done in the aforementioned studies, we note that none of the related studies propose operating on multi-scale volumes which span across all scales of the feature pyramid, nor modifying the label space of the detector to include a localization label for training joint detection and localization models. The experimental analysis section will demonstrate how the detection models benefit from having access to features at different image resolutions in training and testing.

C. Multi-scale structure models (MSS) for object detection

Instead of scoring windows independently across each scale, we propose to operate on the scale volume $\psi(p) = (\phi(p_1), \ldots, \phi(p_S)) \in \mathbb{R}^{d \times S}$ which spans all scales of the feature pyramid. Note that in the MSS approach, the feature pyramid extraction pipeline remains unchanged compared to the baseline. Employing $\psi(p)$ allows for: 1) generalization of the single-scale model approach, 2) Analysis of the role of multi-scale feature-level contextual cues.

The objective of operating on scale-volumes is to resolve the scale label as part of the inference process. Each sample is assigned a label, $y = (y^l, y^b, y^s) \in \mathcal{Y}$ with y^l the object class (in this study, $y^l \in \{-1, 1\}$), $y^b \in \mathbb{R}^4$ is the object bounding box parameters, and y^s is a scale label.

In our experiments, the model dimensions are obtained from the the average box size of all positive instances in the dataset. For each ground truth, y^s is determined by selecting the scale with maximum overlap (area of intersection over union). An overlap of minimum 0.6 is required for a positive sample in any of scale, otherwise the sample is put into the negative set. For instance, for Fig. 1 where the overlap with the ground truth is visualized in red text, $y^s = (00010)$ is the enumeration of the label space.

1) Learning: The multi-scale volume across all scales can be classified into a K class problem, where K is the cardinality of the set of all possible enumerations of y^s. In this work, we set K to be the number of scales S, but in general K may contain more classes than the number of scales.

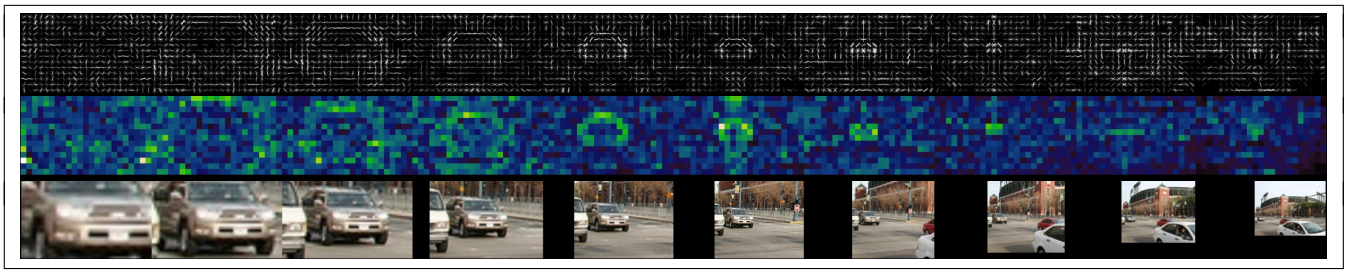
Window scoring in MSS is done using

$$f(p) = \max_{s \in \{1,\ldots,S\}} w_s \cdot \psi(p) \tag{3}$$

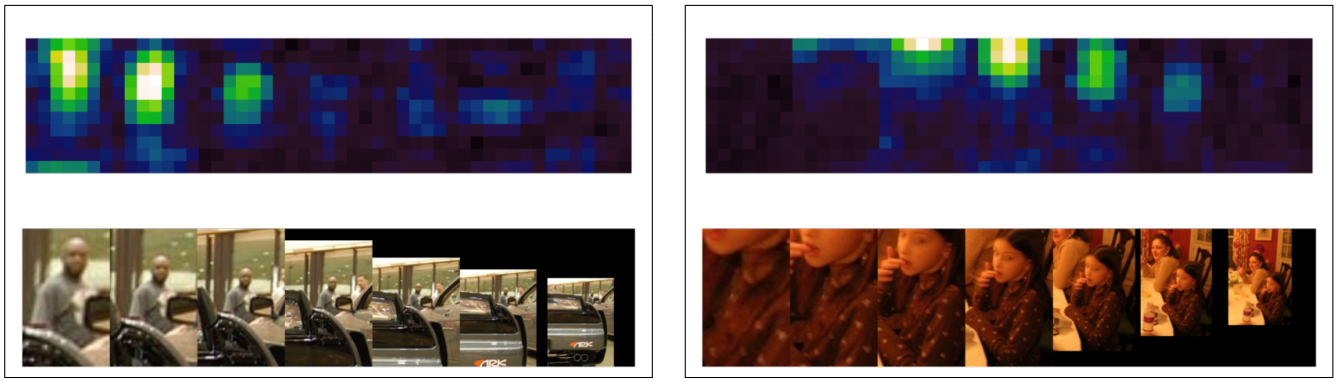
where we learn s model templates, with each spanning the same dimensionality as $\psi(p)$, the multi-scale volume at position p. The same volume is classified into S classes, where the best fitting class is associated with a scale label (obtained with an $\arg\max$ in Eqn. 3). This process predicts the score as well as the final box size at position p.

In order to learn the S linear classifiers parameterized by the weight vectors $w_s \in \mathbb{R}^{d \times S}$, the stochastic dual coordinate ascent solver of [28] with a hinge loss is used. The maximum number of iterations is fixed at 5×10^6 and the tolerance for the stopping criterion at 1×10^{-7} for all of the experiments. Despite significantly higher memory requirements and a large feature vector for the MSS approach, training a single multi-scale template on a CPU takes less than a minute on average. A one-vs-all scheme is used to resolve the best fit scale and score.

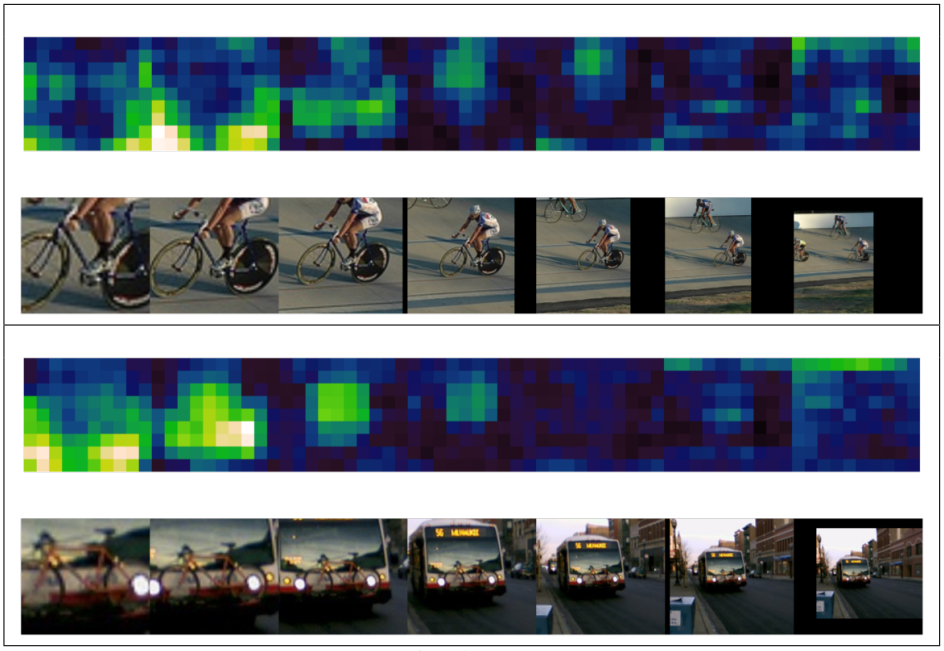
2) Relationship between MSS and the single-scale training baseline: Inspecting Eqn. 3, and comparing it to Eqn. 1, it is clear that Eqn. 1 is a special case of Eqn. 3. For instance, setting w_s to be all zeros outside of the best-fit scale (degenerate case where out-of-scale features are not



Car-HOG



Person-CNN



Bicycle-CNN

Fig. 4: Example MSS templates at a certain scale (w_s in Eqn. 3), depicting the location of discriminative information throughout the volume of multi-scale feature responses for different object categories. We observe how contexual and alignment cues (e.g. road cues for the car object category, rider cues at a different scale for bicycle category, etc.) are selected throughout the multi-scale volume. An example positive sample is visualized for each MSS class.

informative for discrimination of the object class) results in the single-scale model formulation. Under this case, for each level s in the pyramid, $w_s \cdot \psi(p)$ becomes identical to $w \cdot \phi(p_s)$ as in Eqn. 1. We also make the observation that MSS generalizes the template pyramid approach in Eqn. 2, as it learns a scale-specific weight vector w_s. Scale-specific modeling is crucial for robust detection and localization, as the type of contextual appearance cues that the a multi-scale model can capture vary

TABLE I: Detection average precision (%) on VOC 2007 test. Column C shows the number of aspect ratio components. The proposed MSS approach is evaluated with HOG and CNN features.

	C	aero	bike	bird	boat	botl	bus	car	cat	chair	cow	table	dog	horse	mbike	pers	plant	sheep	sofa	train	tv	mAP
HOG	1	13.05	23.54	0.80	1.70	12.85	28.91	27.38	0.68	11.31	8.89	11.04	2.68	13.52	18.49	13.05	5.60	14.58	12.19	16.28	24.48	13.05
HOG-MSS	1	21.72	33.86	10.05	1.81	12.02	22.54	39.80	24.9	13.52	10.08	20.28	13.53	32.57	23.63	23.05	7.24	18.23	22.75	24.20	33.98	20.49
CNN [21]	1	33.54	55.95	24.97	**14.24**	**36.96**	**44.31**	52.33	40.37	**30.07**	44.56	9.09	34.47	51.26	53.39	38.66	**25.22**	40.16	41.36	36.31	**57.97**	38.26
CNN-MSS	1	**41.88**	**56.17**	**30.40**	12.54	25.05	43.36	**60.75**	**50.27**	27.68	**45.41**	**51.25**	**41.94**	**55.60**	**55.71**	**49.30**	22.25	**43.91**	**46.22**	**42.27**	52.78	**42.74**

TABLE II: Detection average precision (%) on VOC 2007 car test set. Column C shows the number of aspect ratio components. Δ AP shows improvement over the baseline used in this work.

method	C	AP	Δ AP
HOG	1	27.38	
HOG	3	32.91	
HOG-MSS (ours)	1	40.04	+12.66
HOG-MSS (ours)	3	49.12	+16.21
CNN $\max_5$ [21]	1	52.33	
CNN $\max_5$ [21]	3	56.90	
CNN-MSS $\max_5$ (ours)	1	60.75	+8.42
CNN-MSS $\max_5$ (ours)	3	**63.23**	+6.33

with respect to the true scale of the object. This can also be seen in the visualization of the learned MSS models (Fig. 4).

III. Experimental Evaluation

The proposed MSS approach is studied on the widely used PASCAL VOC 2007 dataset [5]. For the results, 5 rounds of negative hard mining is performed for all methods, with 5000 negative samples added in each round (beginning with an initial random set which is kept the same for all methods). For HOG, we employ a 10 scale feature pyramid, and for CNN we employ a 7 level pyramid spanning three octaves (scale factor of $2^{-1/2}$ between levels). These were set according to the procedure in Girshick *et al.* [21] in order to perform a fair comparison with a baseline.

Table I demonstrates the results we obtain on the 20 object categories in PASCAL. The overall detection performance improvement is significant for HOG-MSS, by 7.44 mAP points. The improvement due to the proposed HOG-MSS is shown to be directly correlated with the variation in scale within an object class. For instance, classes such as boat, bottle, or potted plant, show a smaller improvement as they contain small scale variation in PASCAL images. On the other hand, classes containing the largest variation in scale, such as cat, train, sofa, dining-table, dog, and horse, show large improvement. Similar trends can be observed for CNN features, which are more scale invariant. CNN-MSS achieves an mAP of 42.74, an improvement of 4.48 mAP points over the results of the publicly available code from [21]. 14 out of the 20 classes exhibit a benefit from incorporating multi-scale cues in training for CNN features, in particular for classes exhibiting large scale variation. Fig. 5 compared the CNN-MSS approach with the baseline in terms of the type of errors made by each algorithm. The comparison is shown on the 'vehicles' superclass of PASCAL (car, motorbike, etc), based on the metrics proposed in [29]. As shown in Fig. 5, incorporation of the MSS framework leads to a significant reduction in localization errors.

As a final experiment, we analyze the improvement due to learning multiple aspect ratio models, either with the baseline or with the MSS framework. Results are shown in Table II, for one and three aspect ratio components. The car object category is used in the experiments as it contains both significant variation in scale and aspect ratio among object instances. The MSS approach is shown to consistently improve car detection performance when increasing the number of aspect ratio components, both with HOG and CNN features. The CNN-MSS three aspect ratio component model reaches 63.23 AP, improving over the single aspect ratio component model by 2.48 accuracy points. Furthermore, CNN-MSS with three aspect ratio components outperforms the R-CNN [7] framework (with features from the same convolutional layer, pool_5), increasing performance from 60.6 to 63.23 AP.

IV. Concluding Remarks

This paper proposed a generalization of the traditional single-scale template detection approach in the aim of better capturing multi-scale information. Training single-scale templates considers features only in a local region. Re-formulation of the problem as a multi-class classification problem allowed the study of a new class of models which were trained to reason over both detection and localization cues. The new set of models significantly improved detection performance when compared to their single-scale template counterparts. In the future, feature selection [30] over scales could significantly reduce the dimensionality of the problem and allow for faster inference run-time (For MSS-CNN, feature extraction is about 0.4 seconds per image with a Titan X GPU and MSS evaluation is about 0.7 seconds per image on a CPU).

V. Acknowledgments

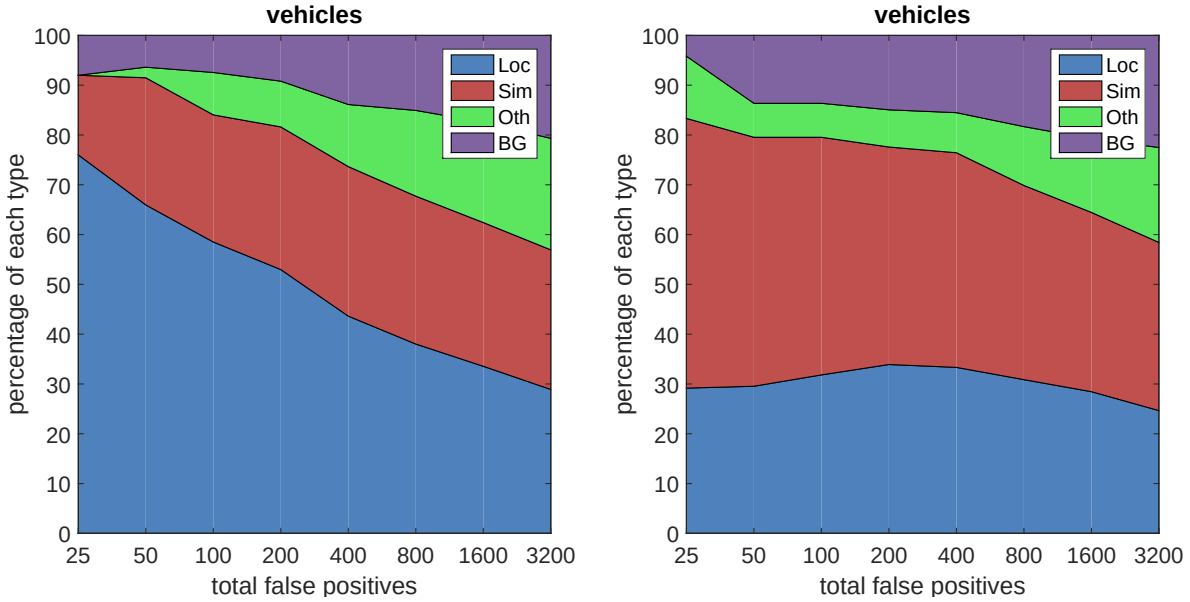
We acknowledge support of associated industry partners and our colleagues at the UCSD CVRR lab for helpful discussions. We also thank the reviewers for their constructive comments.

References

[1] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014. 1

[2] G. Lisanti, S. Karaman, A. D. Bagdanov, and A. D. Bimbo, "Unsupervised scene adaptation for faster multi-scale pedestrian detection," in *ICPR*, 2014. 1

[3] J. Chen, E. Saund, and Y. Wang, "Image objects and multi-scale features for annotation detection," in *ICPR*, 2008. 1



(a) CNN-baseline

(b) CNN-MSS

Fig. 5: Localization quality ('Loc'), as shown for the vehicles type superclass on the PASCAL VOC dataset, is significantly improved when employing the proposed MSS framework. Plots were generated using the error diagnostic tool of [29]. Other types of errors include confusion with similar object categories ('Sim'), confusion with non-similar object categories ('Oth'), and confusion with background ('BG').

[4] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Neural Information Processing Systems*, 2012. 1, 2

[5] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Intl. Journal Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010. 1, 5

[6] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013. 1, 2

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2014. 1, 2, 5

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conf. Computer Vision*, 2014. 1

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010. 1, 2

[10] R. Girshick, "Fast R-CNN," in *IEEE Intl. Conf. on Computer Vision*, 2015. 1

[11] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Intl. Conf. Learning Representations*, 2014. 1, 2

[12] S. Sivaraman and M. M. Trivedi, "Integrated lane and vehicle detection, localization, and tracking: A synergistic approach," *IEEE Trans. Intelligent Transportation Systems*, 2013. 1

[13] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution models for object detection," in *European Conf. Computer Vision*, 2010. 1, 2

[14] S. Sivaraman and M. M. Trivedi, "Vehicle detection by independent parts for urban driver assistance," *IEEE Trans. Intelligent Transportation Systems*, 2013. 1

[15] Y. Ding and J. Xiao, "Contextual boost for pedestrian detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2012. 1

[16] E. Ohn-Bar and M. M. Trivedi, "Learning to detect vehicles by clustering appearance patterns," *IEEE Trans. Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2511–2521, 2015. 1, 2

[17] R. Benenson, M. Mathias, R. Timofte, and L. V. Gool, "Pedestrian detection at 100 frames per second," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2012. 1, 2

[18] R. N. Rajaram, E. Ohn-Bar, and M. M. Trivedi, "Looking at pedestrians at different scales: A multiresolution approach and evaluations," *IEEE Transactions on Intelligent Transportation Systems*, 2016. 1

[19] W. Zhang, G. Zelinsky, and D. Samaras, "Real-time accurate object detection using multiple resolutions," in *IEEE Intl. Conf. on Computer Vision*, 2007. 1

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005. 1

[21] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2015. 1, 2, 5

[22] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Neural Information Processing Systems*, 2014. 2

[23] P. Sermanet and Y. LeCun, "Traffic sign recognition with multi-scale convolutional networks," in *Intl. Joint Conf. Neural Networks*, 2011. 2

[24] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Neural Information Processing Systems*, 2013. 2

[25] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995. 2

[26] R. N. Rajaram, E. Ohn-Bar, and M. M. Trivedi, "Looking at pedestrians at different scales: A multi-resolution approach and evaluations," *IEEE Trans. Intelligent Transportation Systems*, 2016. 2

[27] M. A. Sadeghi and D. Forsyth, "30Hz object detection with DPM V5," in *ECCV*, 2014. 2

[28] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008. 3

[29] D. Hoiem, Y. Chodpathumwan, and Q. Dai, "Diagnosing error in object detectors," in *European Conf. Computer Vision*, 2012. 5, 6

[30] M. Saberian and N. Vasconcelos, "Multi-resolution cascades for multi-class object detection," in *Neural Information Processing Systems*, 2014. 5