# Multi-scale Volumes for Deep Object Detection and Localization

Eshed Ohn-Bar and Mohan M. Trivedi
University of California, San Diego
La Jolla, CA 92093-0434
{eohnbar,mtrivedi}@ucsd.edu

## Abstract

*This study aims to analyze the benefits of improved multi-scale reasoning for object detection and localization with deep convolutional neural networks. To that end, an efficient and general object detection framework which operates on scale volumes of a deep feature pyramid is proposed. In contrast to the proposed approach, most current state-of-the-art object detectors operate on a single-scale in training, while testing involves independent evaluation across scales. One benefit of the proposed approach is in better capturing of multi-scale contextual information, resulting in significant gains in both detection performance and localization quality of objects on the PASCAL VOC dataset and a multi-view highway vehicles dataset. The joint detection and localization scale-specific models are shown to especially benefit detection of challenging object categories which exhibit large scale variation as well as detection of small objects.*

## 1. Introduction

Visual recognition with computer vision has been rapidly improving due to the modern deep Convolutional Neural Network (CNN). The current success is fueled by large datasets, with pre-training of the network for a supervised object classification task on a large dataset [24], and consequent adaptation for new tasks such as object detection [42, 17] or scene analysis [45, 57]. The success of CNNs is attributed to the rich representation power of the deep network. Therefore, much of the current research is concentrated on better understanding properties captured by CNN representations. When transferring the network from a classification task to a detection and localization task, performance is greatly influenced by the ability to capture contextual and multi-scale information [33]. The main aim of this study is in the evaluation and improvement of this ability for CNNs using better multi-scale feature reasoning.

The biological vision system can recognize and locate objects under wide variability due in part to contextual reasoning. This is of particular importance when different image and object scales are considered. Hence, the tasks of capturing contextual cues and modeling multi-scale information are interleaved. Take for instance a car detection task as depicted in Fig. 1. Contextual reasoning appears at different image scales and spatial locations, from fine-grained part information (e.g. bumper, license plate, or tail lights occurring at certain configurations w.r.t. object orientation) and up to contextual scene cues such as road cues or relationship to other objects. Fig. 1 depicts convolutional feature responses computed at twice and half the original image size for a selected feature channel. As can be seen, the responses differ both in magnitude and location depending on the image scale. Responses at different

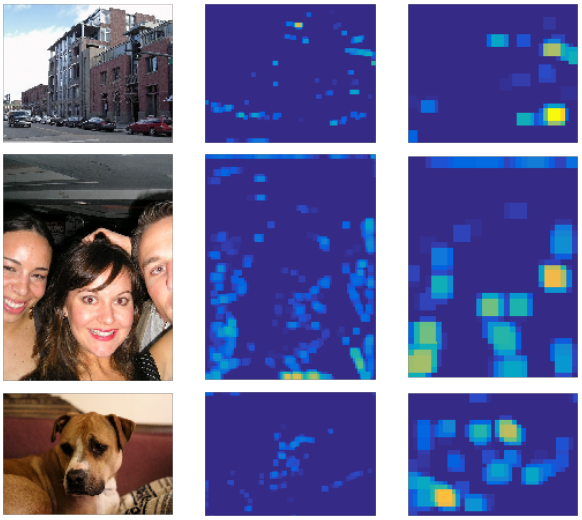


Figure 1: Convolutional feature responses at different image scales of two octaves apart. Different feature channels are visualized for each image. The responses are scale-selective, capturing different levels of contextual information. This phenomenon is studied and modeled in this work using scale volumes in order to obtain better object detection and localization performance.

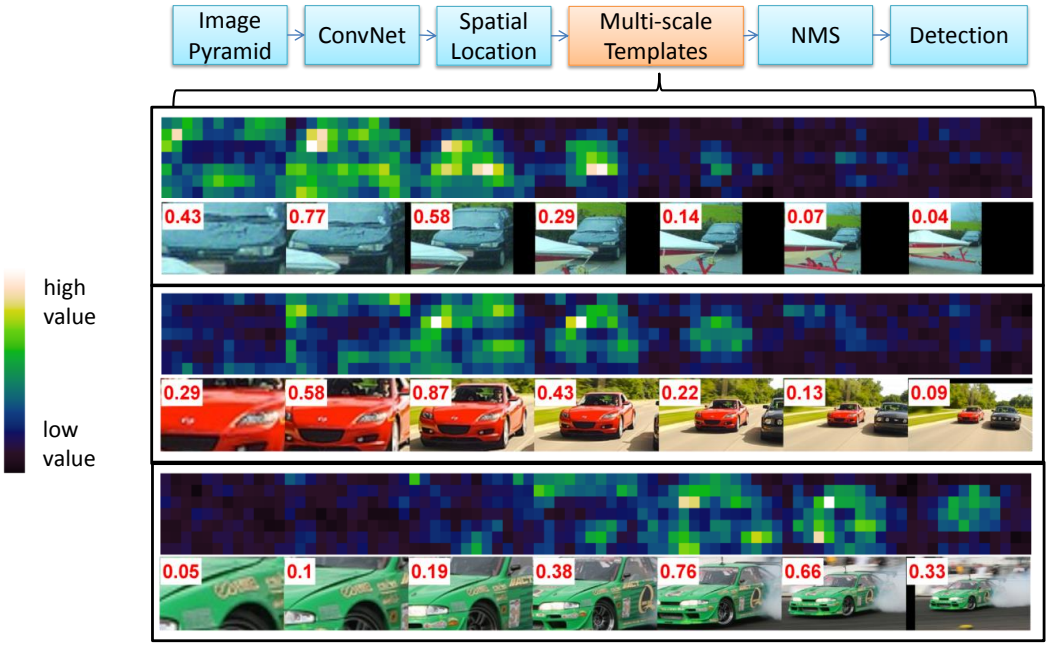



Figure 2: Pipeline of the proposed multi-scale structure (**MSS**) approach for studying the role of contextual and multi-scale cues in object detection and localization. Examples of some of the learned MSS models for 'car' over CNN features are shown, with brighter colors implying greater discriminative value. In red text is the overlap of the annotated ground truth object with a fixed model size. Note how each MSS template selects discriminative information across multiple scales, such as road and part information.

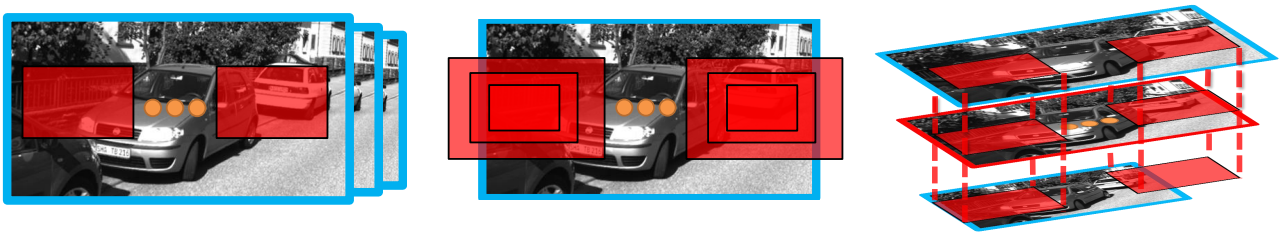
scales contain relevant contextual information for detection and localization. It has been known that CNNs can capture increasingly semantic representations at each layer [54], yet detection performance varies greatly w.r.t. appearance variations (scale, orientation, occlusion, and truncation) [33]. Therefore, contextual multi-scale information can help resolve such challenging cases. This work aims to analyze the benefit of training models that pool features over multiple image scales, both at adjacent and remote scales, on object detection (Fig. 2). Furthermore, the inference label space is adjusted to better leverage contextual multi-scale information in the localization of objects.

### 1.1. Contributions

The main contributions presented in this work are as follows:

1. Multi-scale framework: we propose a framework for understanding CNN responses at multiple image scales. By training models that learn to pool features across multiple scales and appropriately designing the inference label space, the proposed framework is used to perform novel analysis useful in obtaining insight into the role of multi-scale and contextual information. In particular, the impact of dataset size and properties, impact of different scales and object properties, types of detection and localization errors, and model visualization are addressed. The framework generalizes current state-of-the-art object detectors which perform single-scale training and independent model testing across scales.

2. Better detection and localization: Replacing the commonly used local region classification pipeline for detection with a proposed set of joint detection and localization, scale-specific, context-aware, multi-scale volume models is shown to improve detection and localization quality. The contextual information is shown to be particularly useful in resolving challenging objects, such as objects at small scale. Experimental results demonstrate generalization of the proposed, **multi-scale structure (MSS)**, approach across feature types (CNN or hand-designed features) and datasets. The approach is light-weight in memory and computation, and is therefore useful for a variety of application domains requiring a balance between robust object detection and computational cost.



(a) Traditional, single-scale approaches (image pyramid and template pyramid) (b) Our proposed approach

Figure 3: Traditional approaches are limited in ability to capture contextual cues due to a single-scale training and testing of a single-scale local region. The proposed Multi-Scale Structure (MSS) approach extends the local regions across scales of an image pyramid to operate on scale volumes. The inference label space is modified as well to predict a localization label. The access to scale volumes across all scales of the image pyramid in training and testing time allows visualizing contextual cues and analyzing their role in detection and localization.

## 2. Related Research Studies

This study aims to better understand the benefits of improved multi-scale reasoning for object detection and localization. To that end, deep features are extracted at multiple image scales, and models that can perform inference over scale volumes and leverage contextual cues across different scales are trained. The analysis provided by such models is complementary to existing related research studies discussing schemes for object detection and localization with multi-scale, contextual, and deep architectures, as will be discussed below.

**Multi-scale detection**: Traditional multi-scale object detection schemes employ a sliding window, which is a local, fixed-sized region in the image. The local region is scored in a classification task for an object presence, a process done exhaustively over different image locations and at different scales. In training, all training samples are re-sized to a fixed template size, thereby removing any scale-specific information and resulting in a single-scale model. In test time, local regions are classified independently across locations and scales. This limits the model's ability to well-localize an object and capture contextual cues. For instance, the example images in Fig. 2 would be scored independently, despite the highly structured information across scales. Finally, resolving multiple detections is handled with a heuristic Non-Maxima Suppression (NMS) module, which has no access to the image evidence. Several works have challenged this widely used pipeline. This includes the works of [32, 7, 30, 1, 35], which consider training multiple-resolution models. Such techniques were proposed for better handling appearance variation due to scale. The multi-resolution framework of [56] involves rejecting windows at low resolutions before the rest of the image pyramid is processed, thereby achieving speed gains. As the models trained in the aforementioned studies are still single-scale models, testing involves scoring each image location and scale. The contrast between the aforementioned studies and this work is that we incorporate a scale localization label into the label space of the detector, and consequently train models that operate on all scales of the image pyramid (see Fig. 3 for a high-level contrast). The approach explicitly accounts for variation in appearance due to scale and incorporates contextual cues for better localization. We note that the impact on detection and localization quality due to employing features at all scales, both remote and adjacent, has been rarely studied in related studies. As will be shown in Section 3.3, the studied framework generalizes the studies of [32, 7, 30, 1] which do not modify the multi-scale sliding window pipeline, and therefore provides complementary analysis.

**CNN-based object detection**: CNNs are a long-studied class of models [14, 38, 37, 26], achieving impressive performance on a variety of computer vision tasks in recent years [42, 17, 52]. Noteworthy CNN-based detection schemes are the OverFeat [42] and Region-based CNN (R-CNN [17]) architectures. Although both employ a CNN, OverFeat performs sliding-window detection (which is common in traditional object detection), while R-CNN operates on a set of region proposals. We note that both [42, 17] operate in a local-region manner without joint reasoning over multiple scales of an image pyramid. Current improvements over such architectures emphasize 1) The learning and incorporation of deeper networks [4, 46], 2) Resolving different components of the successful R-CNN framework into a single, end-to-end architecture. The original R-CNN framework involves a multi-stage pipeline, from object proposal generation (e.g. Selective Search [49]) to SVM training and bounding box regression. At test-time, a CNN forward pass is performed for each region proposal, which is costly. In contrast, SPPnet [19], Fast R-CNN [16], and OverFeat require only a single forward pass. Fast R-CNN [16] employs a Region of Interest (ROI) pooling layer which operates on region proposals projected to

the convolutional feature map. Furthermore, the bounding box regression module is also integrated into the end-to-end training using a sibling output layer. Recently, another boost in performance was introduced in Faster R-CNN [36], which incorporates a Region Proposal Network in order to improve over the Selective Search region-proposal module. Independent testing at multiple scales is shown to improve performance on the PASCAL benchmark in the aforementioned studies, yet no further analysis is shown. Larger gains from multi-scale analysis are generally shown for other domains requiring robustness over large scale variations such as on-road vehicle detection [15] and pedestrian detection on the Caltech benchmark [53, 43]. In general, common CNN and hand-crafted object detectors involve training for and classifying a local region with a single-scale model. The contextual modeling capacity of such models is therefore limited, and detection of objects at multiple scales is done by independent scoring of an image pyramid. Nonetheless, visual information across scales at a given image location is highly correlated. Therefore, pooling features over scales in training and testing may benefit an object detector. Our work leverages a novel multi-scale detection framework in order to study the role of contextual information across image scales in a given spatial location.

**Contextual object detection**: Our study is relevant to the study of context. Classifying scale volumes directly benefits from contextual cues found at different levels of an image pyramid. Hence, scale and context modeling are interleaved fundamental tasks in computer vision [31, 6, 52, 10, 20]. Careful reasoning over these two tasks has shown great success in a variety of computer vision domains, from image segmentation [12] to edge detection [52]. The Deformable Part Model (DPM) [13, 55] is another example, as it reasons over a lower resolution root and higher resolution parts templates. Commonly, an additional module for capturing spatial and scale contextual interactions is applied over the score pyramid output of a traditional local-region, single-scale detector [17, 28, 13, 22]. In contrast, the studied framework in this work joins the two steps. In Chen *et al.* [5], a Multi-Order Contextual co-Occurrence (MOCO) framework was proposed, extending the Auto-Context idea [48, 34] for context modeling among boxes produced by traditional local region detection schemes. Sadeghi and Farhadi [39] propose visual phrases to reason over the output of object detectors and local context of object relationships. Desai *et al.* [6] formulate multi-class object recognition as a structured prediction task, rescoring object boxes and replacing NMS for improved modeling of spatial co-occurrence. Li *et al.* [27] propose a hierarchical And-Or model for modeling context, parts, and spatial arrangements, and show large detection performance gains at a car detection task. Unlike the aforementioned, this work aims to study the benefit of incorporation of contextual, multi-scale cues directly into to object detection scheme. This is done both by modifying the detector to operate on scale volumes spanning the entire image pyramid and the inference label space. Analysis regarding the impact of such a framework is lacking in the aforementioned studies.

**Multi-scale deep networks for contextual reasoning**: Multi-scale deep networks have been previously studied in [10, 12, 44]. Eigen *et al.* [10] predicts depth maps by employing two deep network stacks, one for making coarse global prediction over the entire image and another for local refinement. Similarly to [10], this work aims to analyze the role of capturing information at different image scales. In contrast to [10], we discus the task of object detection and localization, study deep features at more than two image scales, and aim to better capture image appearance variations due to scale. Sermanet *et al.* [44] propose a multi-scale branched CNN for traffic sign recognition. Here, scale refers to different levels of feature abstraction as opposed to image pyramid scales. Although related to our study in capturing context, the method does not employ feature responses or weight learning across image scales for handling scale variation and improved object localization.

A close approach to ours is the work of Farabet *et al.* [12], which proposes a multi-scale CNN for semantic scene labeling of pixels. Consequently, segmentation quality is significantly improved by learning CNN weights which are shared across three image scales. Commonly, multi-scale architectures employ 2-3 image scales at most, while we employ 7-10, and modify the inference label space. The multi-scale CNN is shown in [12] to be better at capturing image evidence at a certain pixel location, yet no insight is given regarding th impact at different object scales (e.g. small objects), contribution of weights at different scales, relationship between object class and context usefulness, or impact on localization quality. Generally, adding responses at multiple image scales is known to benefit a variety of vision tasks, yet analysis on its role for general object detection and localization is lacking. Our study is also motivated by the fact that most current state-of-the-art object detectors do not employ multi-scale features or modeling [42, 17, 16, 36]. Furthermore, the training formulation in this work allows for visualization of the multi-scale, contextual cues. In contrast, most related studies discuss improvement due to multi-scale image features on a performance level only (e.g. features with one image scale vs. two image scales), without providing further insights.

## 3. Capturing Context with the Multi-Scale Structure (MSS) Approach

The main approach in which context in object detection will be studied is presented in this section. The method is contrasted with existing schemes which are limited in their

contextual reasoning in Fig. 3. Instead of training and testing over local image regions (either a sliding window or region proposals), the approach employs an image pyramid and operates on features at all scales in training and testing. As large scales include fine-grained information, such as part-level information, and small scales include scene-level information, the MSS approach allows a study of the importance of cues at different scales. Furthermore, scale-specific multi-scale models are trained as contextual cues vary greatly w.r.t. the object scale. The MSS approach is also directly comparable to traditional single-scale training/testing baseline as the feature pyramid input to both is kept the unchanged.

### 3.1. Efficient Feature Pyramids

In order to efficiently train and test models which reason and pool over multi-scale features, all experiments are performed in an architecture similar to OverFeat [42, 47] and DeepPyramid DPM [18]. These have shown powerful generalization and flexibility to a variety of tasks, even without fine-tuning [53]. Hence, they are suitable for studying the ability to model context when transferring from the ImageNet classification task to the detection task. Furthermore, they provide *simple and efficient* means for handling multi-scale image pyramid information (order of magnitude faster than the original and widely used R-CNN [17]). By only employing the convolutional layers (discarding the fully connected layers), spatial structure is preserved and image regions can be directly projected to feature responses in an efficient manner without requiring a region proposal mechanism. Although more intricate approaches exist which preserve the fully connected layers (such as faster R-CNN [36]), the used ROI pooling layer in existing approaches still *operates on a single scale* of image features, and so the approach is orthogonal to our study. The network we employ is a truncated version of the winning network of the ILSVRC-2012 ImageNet challenge [24] composed of 8 layers in total. The network is used as a main tool to better understand context in CNNs. Employing deeper networks [36, 46] greatly improves performance by improving *local classification* power, but these are generally evaluated in a single-scale manner (or independent evaluation over multiple scales) and so are also orthogonal to this study. As tasks with large scale variation (e.g. pedestrian detection [53, 8]) require a large image pyramid in order to reach state-of-the-art performance, the approach in this work is also motivated by the need of real-world applications for a trade-off between performance, computational efficiency, and memory requirements. Our study of efficient multi-scale contextual reasoning is directly applicable to such applications.

### 3.2. Multi-scale detection with a single-scale template

First, we introduce notation to clarify and motivate the MSS approach. In traditional object detection, context reasoning is limited as detection is performed in a single scale fashion (tested independently at multiple image scales). First, a feature pyramid is constructed over the entire image at each scale to avoid redundant computation for each striding window. 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. Most of the analysis will involve a single aspect ratio model (which is common), and so we do not include that additional parameter in $p_s$, yet the formulation supports multiple aspect ratio models [29]. Generally, the feature pyramid is at a lower spatial resolution than that of the image of the same scale (due to convolution and subsampling). 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. Mapping locations over scales can be achieved by a multiplication by the scale factor as well. Each window contains an array of feature values, $\phi(p_s) \in \mathbb{R}^d$, to be scored using a filter $w$ learned by a discriminative classifier, in our case a support vector machine (SVM). The scoring is done using a dot product,

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

Generally, the template size is defined as the smallest object size to be detected, and further reduction in template size results in degradation of the detection performance. Note that learning and classification only occurs over a local window. A similar pipeline can be described using a template pyramid as studied in [1, 40, 32] and was shown to improve results due to capturing finer features at different scales that would have been discarded by the downsampling. In this approach, a set of templates are learned, $(w_1, \ldots, w_S)$. In detection, the $S$ templates are evaluated so that each location $p$ in the original image scale is scored using the set of model templates

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

where we drop $s$ as only one scale of the image is considered. We emphasize that the model filters in this approach are also trained on locally windowed features only, but may capture different cues for each scale. In principle, this approach is similar to the baseline as it performs the scoring convolution at each scale independently of all other scales (unlike MSS, as shown in Fig. 4).

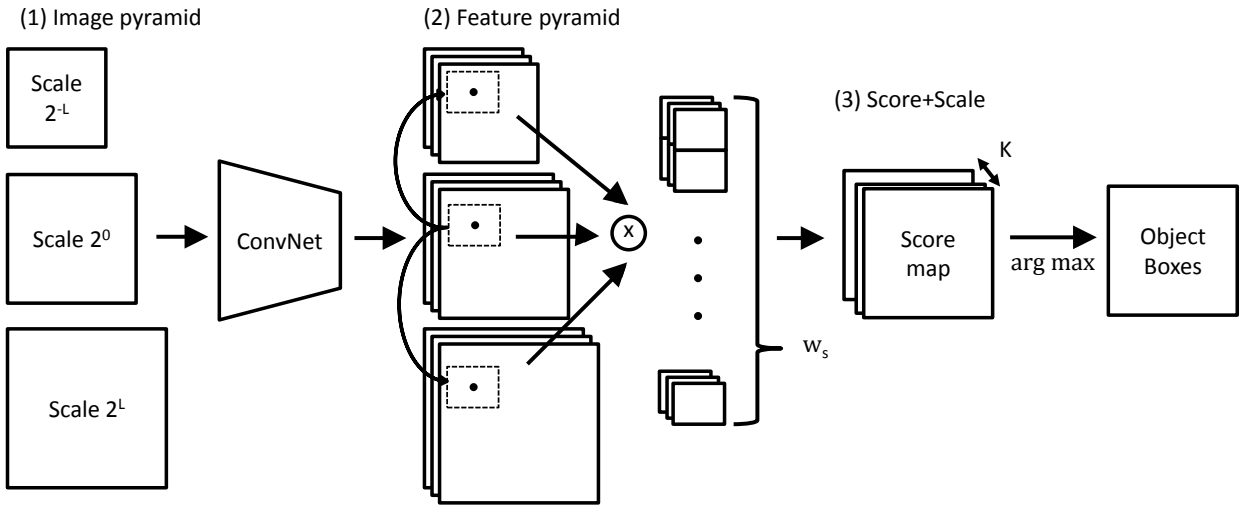



Figure 4: Our proposed approach re-samples the original image to obtain an image pyramid. Object-level annotations are converted to multi-scale annotations by obtaining a scale label. The scale label is assigned for each sample based on an overlap of the ground truth in each scale with a fixed model size (Section 3.3.1). Each sample is associated with a feature array that is cropped from the feature pyramid at shifted versions for preserving the same spatial location across scale. Testing involves scoring (represented by the 'X' operation in the figure) using learned multi-scale templates which convert the feature pyramid to an object score map. Note that the feature maps for each scale shown in the figure are at a lower spatial resolution than the original images.

## 3.3. Multi-scale detection with a multi-scale template

The feature pyramid computation and handling is mostly left unchanged in the proposed MSS approach. Spatial locations in the image space can be mapped across scales using a scale factor. As shown in Fig. 2, evaluations at the same spatial location occur repeatedly over scales. This mechanism is replaced by considering features from all scales at a given image location, i.e. $\psi(p) = (\phi(p_1), \ldots, \phi(p_S)) \in \mathbb{R}^{d \times S}$ descriptor.

### 3.3.1 Label space

Next the process of labeling training samples is outlined. 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 only $y^l \in \{-1, 1\}$ is considered), $y^b \in \mathbb{R}^4$ is the object bounding box parameters, and $y^s$ is a scale label. In our experiments, the model dimensions are obtained from the average box size of all positive instances in the dataset (providing a single aspect ratio model). Training instances are sampled directly from the feature pyramid in a simple process where, 1) the multi-scale template is centered on top of each ground truth window spatial location and 2) Overlap with the ground truth is checked in each image scale (as shown in red in Fig. 2). Formally, a vector of overlaps $F$ is constructed. If the image at $s$-th level contains $\hat{y}(s) = \{\hat{y}_1(s), \ldots, \hat{y}_N(s)\}$ ground truth boxes, the template box is centered on a positive sample at the $s$-th level (denoted as $B(s)$), so that entries of $F$ are computed for each pyramid level,

$$F(s) = \max_{i \in \{1,\ldots,N\}} \text{ov}(B(s), \hat{y}_i^b(s)). \quad (3)$$

where ov($a$,$b$) = area($a \cap b$)/area($a \cup b$) for two rectangles, $a$ and $b$. $F$ is shown for three examples in Fig. 2. For instance, for Fig. 2 first row, $y^s = (0100000)$. Peaks in $F(s)$ with high overlap imply a positive instance. This process potentially allows for multiple labels over scales to be predicted jointly, i.e. two almost overlapping objects at different scales, but such instances are rare. For simplicity, we only allow a single scale-label association by employing the scale where maximum overlap occurs.

### 3.3.2 Learning

Two max-margin approaches are studied for learning the multi-scale object templates, leveraging the highly structured multi-scale information, and analyzing importance of contextual information at different scales. Such information would have been ignored if a single-scale template was used.

Parameterization in the image pyramid can be done once over spatial locations at different scales by mapping across region locations with a scaling factor. Although these lo-

cal regions across scales remain the same both in a traditional single-scale model classification procedure and the MSS approach, this new parameterization implies that we can concatenate features at all scales, as opposed to classifying these separately across scales. Furthermore, the previous section showed how such samples could be labeled, so the problem can now be posed as a multi-class problem.

**One-vs-All**: There are well developed machine learning tools for dealing with a large-dimensional multi-class classification problem. A straightforward solution is with a one-vs-all (OVA) SVM, which allows training the multi-class templates quickly and in parallel. Window scoring is done using

$$f(p) = \max_{s \in \{1,\ldots,K\}} w_s \cdot \psi(p) \tag{4}$$

The scale of the box is obtained with an $\arg\max$ in Eqn. 4. In order to learn the $K$ linear classifiers parameterized by the weight vectors $w_s \in \mathbb{R}^{d \times S}$, the stochastic dual coordinate ascent solver of [50] with a hinge loss is used. The maximum number of iterations is fixed at $5 \times 10^6$ and the tolerance for the stopping criterion at $1 \times 10^{-7}$ for all of the experiments. Training a single multi-scale template on a CPU on average takes less than a minute.

For simplicity, this paper considers training a model for each scale, so that $K = S$. In general, this may not be the case (e.g. pedestrians occurring at close proximity but at different scales).

**Structured SVM**: A second approach can be used in order to learn all of the multi-scale templates jointly. A feature map is constructed using the labels of each sample as following,

$$\Phi(p, y) = (\Psi_1(p, y), \ldots, \Psi_K(p, y)). \tag{5}$$

$$\Psi_k(p, y) = \begin{cases} \psi(p) & \text{if } y = k \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

This approach allows for learning a joint weight vector over all classes $w = (w_1, \ldots, w_K)$, such that

$$f(p) = \max_{y \in \mathcal{Y}} w \cdot \Phi(p, y) \tag{7}$$

Where the scale label prediction similar to as in Eqn. 4, but the loss function in training is defined differently using other elements of $y$.

Given a set of image-label pairs of the form $\{p^i, y_i\}$, the model is trained using a cost-sensitive SVM objective function [23, 3, 25]

$$\begin{aligned} &\min_{w,\xi} \frac{1}{2}||w||^2 + C \sum_{i=1}^{n} \xi_i \\ &\text{s.t. for } \forall i, \bar{y} \in \mathcal{Y} \setminus y_i \\ &w \cdot (\Phi(p^i, y_i) - \Phi(p^i, \bar{y})) \geq L(y_i, \bar{y}) - \xi_i \end{aligned} \tag{8}$$

The loss function, $L$, is chosen to favor large overlap with the ground truth,

$$L(y, \hat{y}) = \begin{cases} 0 & \text{if } y^l = \hat{y}^l = -1 \text{ or} \\ & \max\limits_{i \in \{1,\ldots,N\}} \text{ov}(y^b, \hat{y}_i^b) < 0.6 \\ 1 & \text{otherwise} \end{cases} \tag{9}$$

#### 3.3.3 Generalization of the single-scale approach

The main aim is to study context. The purpose of introducing the MSS approach is that it generalizes the traditional single-scale approach. Below, we show that in principle, if other scales do not contain additional contextual information, MSS reduces to the traditional single-scale approach. Eqns. 4 and 7 employ features at all scales for a given spatial location. Such a formulation allows learning the class weights jointly, as in Eqn. 7. It can be shown that this is a generalization of the single-scale template baseline. For instance, if no discriminative value is added by adding features at different scales, then the corresponding weights $w_s$ in Eqn. 4 will only select features in the single best-fit scale (i.e. a degenerate case). Therefore, for each level $s$ in the pyramid, $w_s \cdot \psi(p)$ becomes identical to $w \cdot \phi(p_s)$ as in Eqn. 1. A similar argument demonstrates the same for Eqn. 7. Therefore, both of the studied multi-scale template learning approaches can benefit by having access to additional information not accessible to the single-scale template approaches which only employs local window features at one scale. Furthermore, by learning a separate weight for each class, the model can account for appearance variations at different resolutions [1] and learn scale-specific context cues.

## 4. Experimental Evaluation

The experiments aim to quantify the importance of context cues in deeply learned features for a detection and localization task. Initially, the MSS approach is developed on the PASCAL VOC 2007 dataset [11] using its established metrics, followed by analysis on a multi-view highway vehicles dataset with large variation in object scale.

**Features**: Two representative visual descriptors are employed in order the study the role of context. Most of the experiments involve the deeply learned features discussed in Sec 3.1. The fifth convolution layer output has 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 Girshick *et al.* [18], the CNN features already provide part and scale selective cues. This can be enhanced by applying a $3 \times 3$ max-pooling layer. For

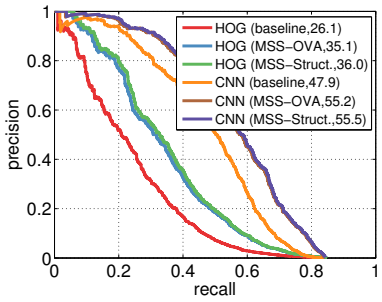



Figure 5: Model training comparison on a validation set for 'car' detection using HOG and $\text{conv}_5$ features. Average Precision (AP) is shown in parenthesis. Contextual information captured with MSS is shown to significantly improve detection performance using both one-vs-all (OVA) and structural SVM (Struct.) training.

direct comparison with the DeepPyramid approach [18], the same feature extraction and 7-scale pyramid pipeline was implemented in the experiment. The HOG feature implementation of [13] serves as a comparative baseline and studying generalization of experimental analysis across different feature types. HOG is used with a cell size/stride of 8.

**Image pyramid**: The scale factor between levels is set to $2^{-1/2}$. The CNN feature pyramid spans three octaves with 7 levels. For HOG features, adding 3 more levels to the image pyramid for a total of 10 was shown to improve performance. In all of the experiments, training instances are extracted directly from the feature pyramid, as opposed to extracting features from cropped image samples. For the CNN feature pyramid, the features used are computed by the fifth convolutional layer which has a large receptive field of size $163 \times 163$ pixels.

**Data augmentation**: Training images are scale-jittered by up to an octave (either down-sampled and zero-padded or up-sampled and center-cropped). In addition to flipping, this data augmentation was essential for obtaining good performance of the MSS approach on all of the object categories.

**Hard negative mining**: All approaches studied employ an iterative process by which hard negatives are collected for re-training. The process eventually converges, when the number of negative samples generated are below a certain threshold. All of the experiments begin with a random set of 5000 negative samples. For a given object category, the initial negative samples are kept the same across techniques to allow direct comparison. In each iteration, up to 5000 additional negatives are collected. For mining, both images containing positive instances and negative images are used. A threshold of $0.3$ overlap is used for mining negative samples from images with object instances.

### 4.1. Analysis on the PASCAL VOC dataset

**Learning framework choice**: First, we evaluated the choice of learning framework on a validation set of the 'car' category. Fig. 5 details the analysis of different learning and features combinations on the car category. Context is shown to benefit both HOG and $\text{conv}_5$ CNN features, as both learned MSS detectors are shown to greatly outperform the baseline in detection Average Precision (AP). Training the templates using the structural SVM allows for joint learning of the MSS templates, yet the improvement is marginal. Because structural SVM training is more costly, one-vs-all models are employed for the remainder of the experiments in this study. The structural SVM formulation may be of interest in the future for bounding box regression [2] or parts integration [13].

**Visualization of the learned models**: Fig. 6 depicts some of the learned MSS models for different object categories (positive valued entries in a learned MSS weight model). A single multi-scale template is visualized with a corresponding positive instance for each object category. For a given spatial location in the model, we visualize the learned model weights at each scale. As shown, while the best-fit scale includes large amount of the discriminative value, features from other scales (both adjacent and remote) are also selected. Contextual patterns can be seen, such as selection of road cues for car detection. We also observe the existence of alignment features, where certain appearance cues at one scale may assist in localization at another scale. This is shown by a repetitive shape pattern across the scales.

**Relationship between scale-variation, dataset size, and MSS benefit**: Our experiments showed the MSS method to significantly impact performance on some object classes by up to 7 AP points (e.g. 'bottle' and 'dining table' classes). Overall, 12 out of the 20 object categories benefit from the MSS approach, specifically on challenging object instances (i.e. small objects) and in terms of localization quality. Furthermore, overall mAP is improved with the MSS approach as shown in Table 1. Nonetheless, certain object categories do not benefit from incorporation of the multi-scale reasoning. As the reason for this is not immediately clear, we further study it next. A closer inspection of the scale distribution of the different classes reveals some insight, as shown in Fig. 7. First, a difference between HOG and CNN features is observed. Because CNN features are more scale-sensitive than HOG, this translates into smaller performance gains due to multi-scale context. Employing HOG on the other hand results in large gains consistently and across all object categories. A second observation is that some classes in the PASCAL VOC dataset exhibit smaller variation in scale. This limits the benefits due to incorporation of multi-scale context, and results in smaller AP improvement. If a certain object class exhibits

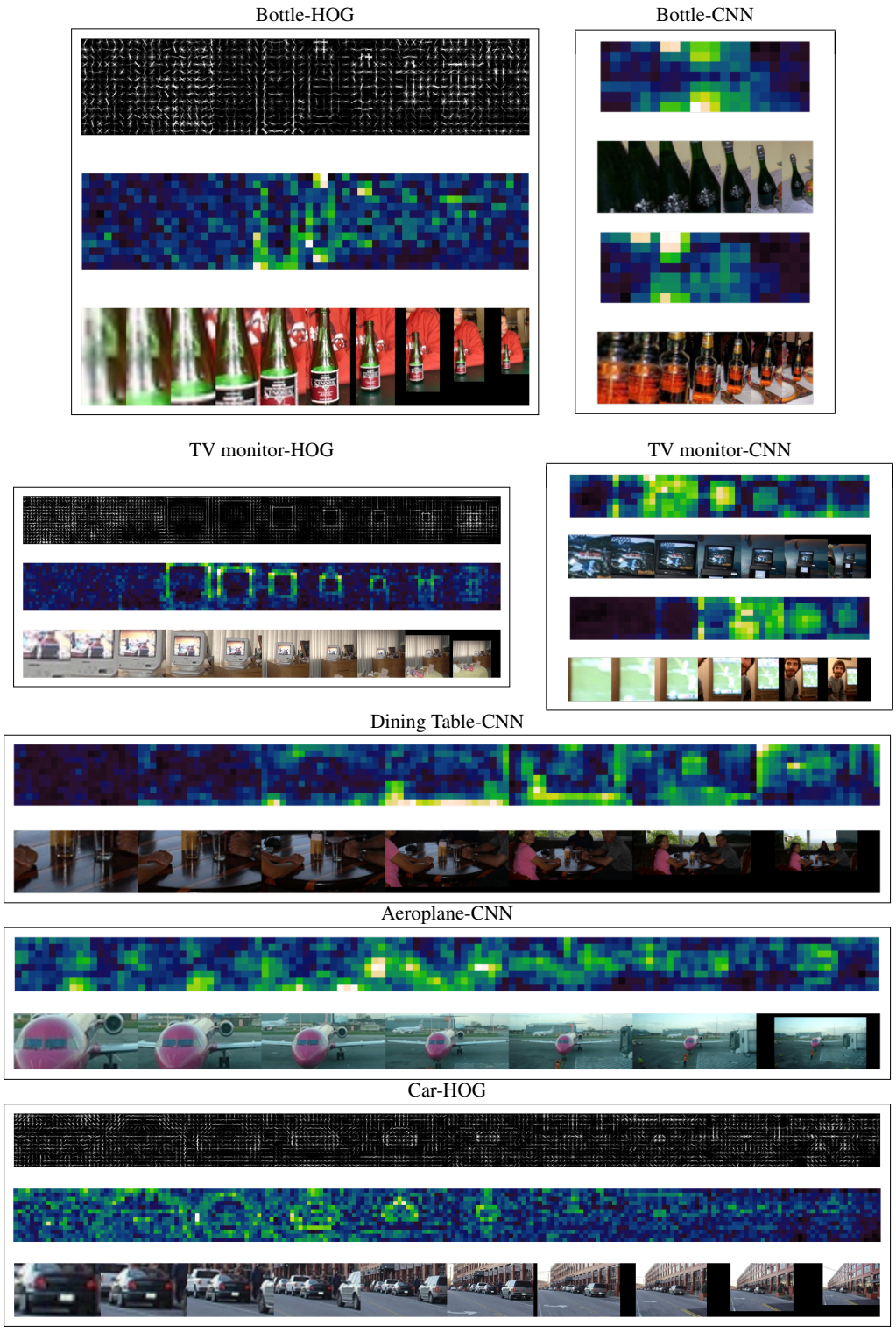
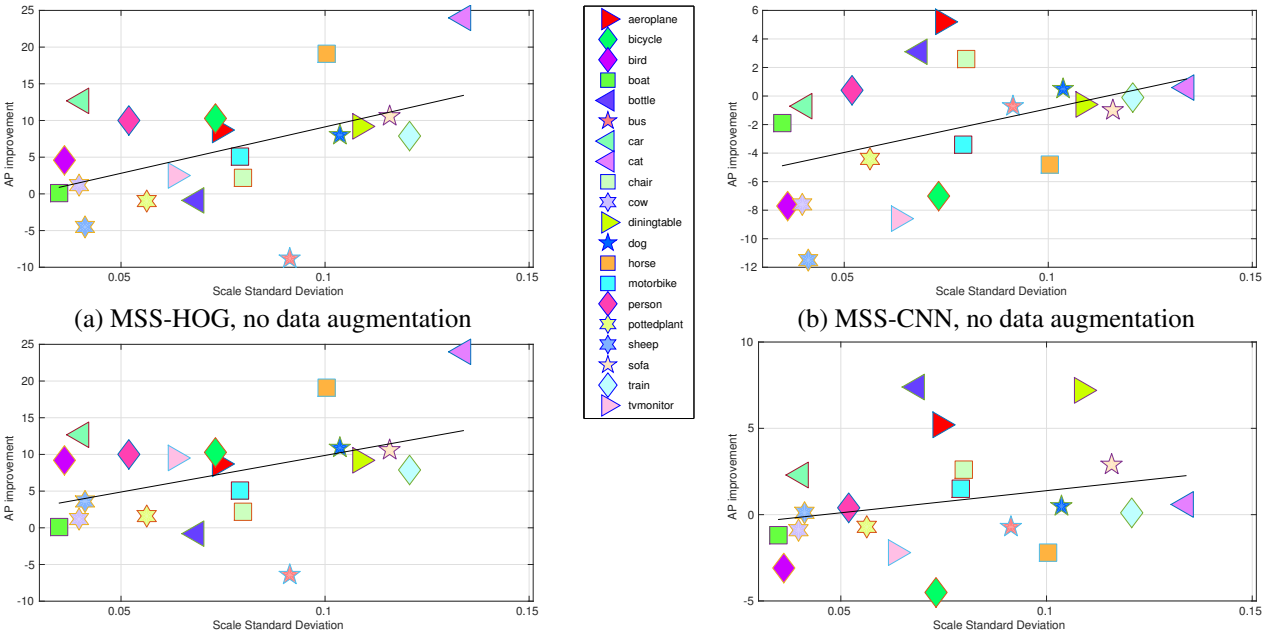



Figure 6: Visualization of multi-scale CNN and HOG templates. For each model, the maximum positive SVM weight for each block is shown together with an example instance. Brighter colors imply higher discriminative value. Large amount of discriminative value is placed at nearby and remote scales corresponding to contextual information (e.g. road cues at other scales).



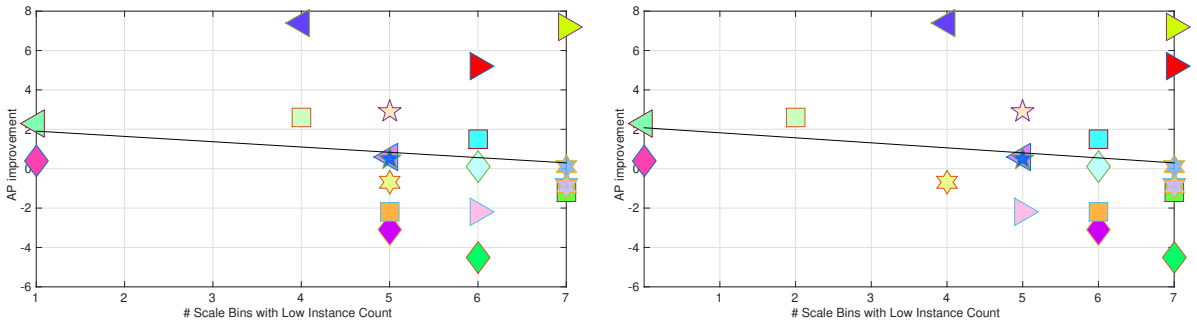

(a) MSS-HOG, no data augmentation

(b) MSS-CNN, no data augmentation

(c) MSS-HOG, with data augmentation

(d) MSS-CNN, with data augmentation

Figure 7: Relationship between the scale distribution of class samples in test time and the corresponding improvement in AP with the proposed MSS approach. As shown, our method shines when there is a large spread in the distribution over scales. Although some classes tend to appear in the PASCAL VOC dataset in a narrow scale distribution, this phenomenon is dataset and object specific. Therefore, if more instances at varying scales were to be added, the proposed approach would be better suited for such settings.

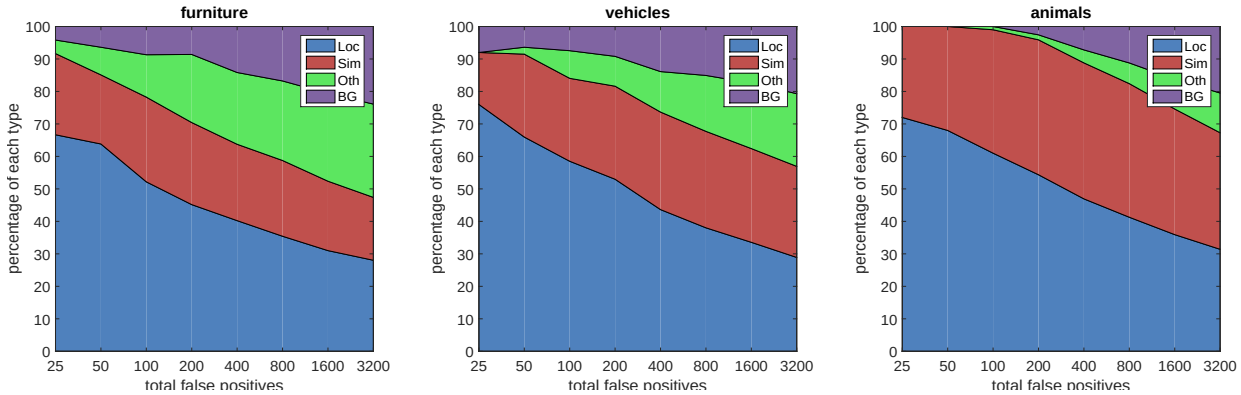



(a) Training set

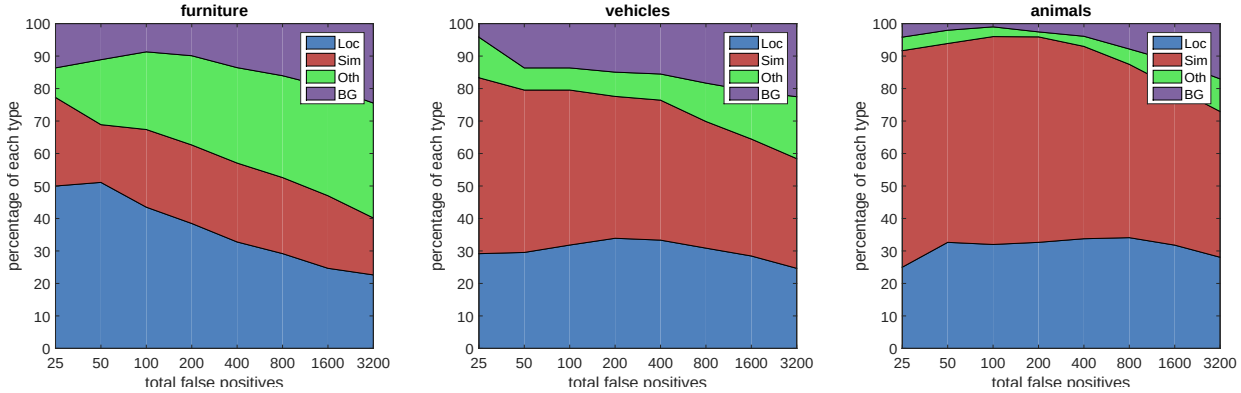
(b) Testing set

Figure 8: Relationship between dataset properties and performance of the CNN-MSS approach. Some of the object classes in the PASCAL VOC benchmark contain a small number of object instances at multiple object scales, which poses a challenge to the scale-specific MSS models.

smaller scale variation in the test set, the contextual cues will be less beneficial, which implies the results are influenced by the object statistics in the test set. Finally, we wish to analyze the role of dataset size on the variation in performance. Because the multi-scale templates require scale-specific instances, a small number of instances in the dataset (even with data augmentation) could lead to sub-optimal learning and consequent reduction in performance gains. The importance of sufficient training instances for training each of the scale-specific MSS template is verified in Fig. 8. As shown in Fig. 8, classes with low detection AP improvement also contain a small number of objects in multiple image scales. In Fig. 8, low instance count is defined as a value under the average number of instances per scale bin across all object categories. Together with the observations in Fig. 7 regarding limited scale variability and insufficient training data explain why detection of certain classes, such as 'bottle', 'aeroplane', 'dining-table', and




(a) CNN-baseline




(b) CNN-MSS

Figure 9: Analysis of the distribution of false positive types [21] for different types of objects on PASCAL VOC 2007. Training and testing is done with a single aspect ratio model. Loc - poor localization, Sim - confusion with a similar category, Oth - confusion with non-similar object category, and BG - confusion with background. The MSS approach is shown to significantly reduce errors due to poor localization.

'sofa', greatly benefit from the multi-scale context framework, and some classes do not (mainly 'boat' and 'bird' which contain small scale variability as shown in Fig. 7)). As will be shown next, the MSS approach significantly improves localization quality across all object categories.

**Localization quality**: Fig. 9 demonstrates improved localization due to incorporation of contextual cues across scales. The improvement is consistent over all types of object categories (clustered into three super-classes), including furniture, vehicles, and animals. This type of analysis is encouraging, as CNN-based object detectors are known to suffer from in-accurate localization. Our approach demonstrates the benefit on localization due to explicit incorporation of multi-scale features. This is intuitive, as the existence of certain feature responses at some scales can assist in better localization at another scale.

**Context statistics**: Training MSS models places discriminative value on each multi-scale cue. Next, we aim to understand how important are such cues in the learning process. For each class, features were divided into two: 1) Features found in the best-fit scale corresponding to the same features that would be employed if a single-scale template (referred to as 'in-scale' features), and 2) 'out-of-scale' features which a placed outside of the best-fit scale. The learned parameters, $w$, can be decomposed to positive and negative valued entries as $w = w^+ + w^-$. Indices with higher absolute value correspond to locations in the feature space which provide large discriminative value. Single-scale model training involves only 'in-scale' features. Furthermore, if 'out-of-scale' features provided no benefit, we would expect the majority of the discriminative weight to be placed on the best-fit 'in-scale' features only.

By studying the percentage of discriminative weight in $w^+$ and its distribution across scales for MSS-CNN, Fig. 10 demonstrates the clear trend of choosing features that are placed outside of the ground truth scale in training. This is a data-driven affirmation of the proposed approach. Although only positive weights shown in Fig. 10, the trends are simi-

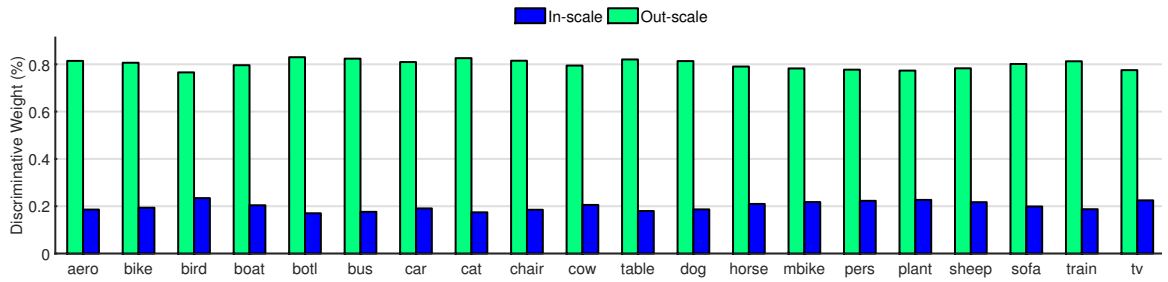



Figure 10: For CNN-based detection at a given scale, how important are out-of-scale context features? See Sec. 4.1 for details.

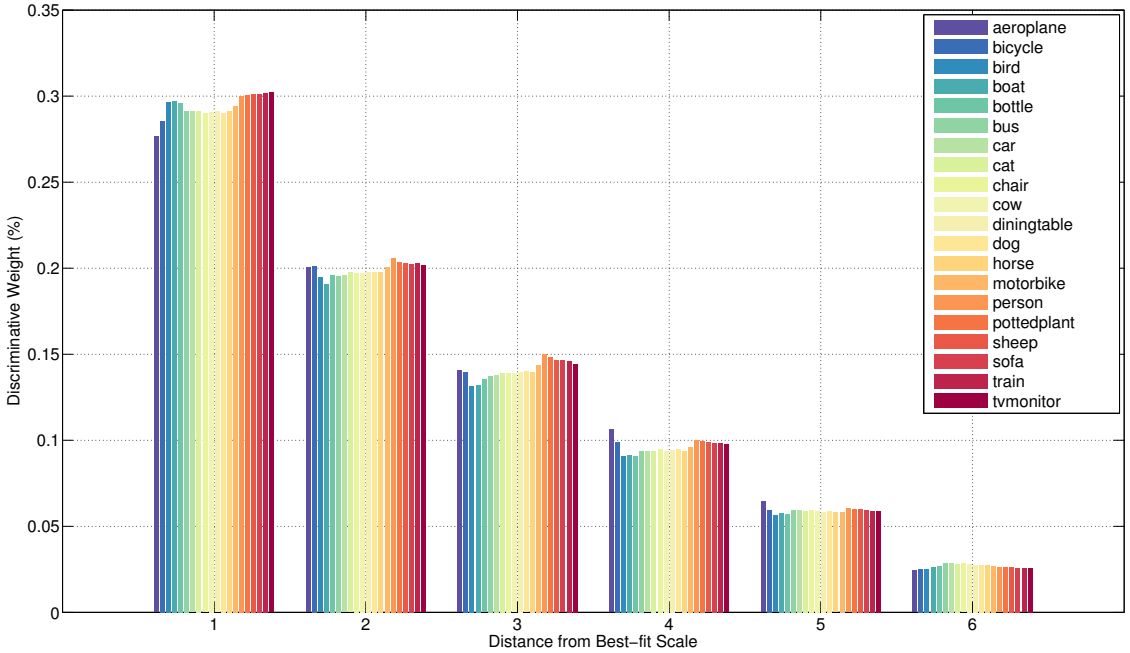



Figure 11: Relative to the best-fit scale, how is discriminative value distributed across pyramid levels? Most of the weight is found within adjacent levels (distance of '1' level away), but the contextual cues are shown to span all levels.

lar both over positive weights $w^+$ and negative weights $w^-$. We can see that context can benefit CNN-detection greatly.

A further breakdown of this information is visualized in Fig. 11. Here, it is shown that most of the features selected outside of the best-fit scale are located in the adjacent scale (a distance of '1' pyramid level away), which is to be expected. Nonetheless, the MSS models consistently select features at more remote pyramid levels, even up to more than an octave away. This analysis suggests that CNN-based approaches can greatly benefit from careful multi-scale and contextual reasoning, which is not done in most existing approaches for object detection. Simple pooling over both adjacent and remote scales is shown to greatly assist in detection, as shown in Fig. 11. Interestingly, a spike at certain remote scales is clearly seen with some categories, such as 'aeroplane', 'bicycle', and 'person'. This observation can be better understood by inspecting the template visualization in Fig. 6. For 'aeroplane', many of the scales contain informative contextual information as shown in Fig. 6, from wings to other aeroplanes. For 'bicycle', a rider may be found at a further scale. It can also be clearly observed how classes which MSS benefits least ('bird' and 'boat') have the smallest discriminative value placed in other scales out of all object categories. In these classes, contextual information is not selected as much.

**Performance breakdown by scale**: As shown in Fig. 12, most gains in detection performance with CNN-features come from detection of smaller objects (50 pixels and less in height). This is intuitive, as such objects can benefit from incorporation of contextual cues at other scales.

| | 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 | 40.04 | 24.66 | 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 [18] | 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-ours | 1 | 36.68 | **60.66** | **33.45** | 13.71 | 17.66 | 44.02 | 58.48 | 49.71 | 25.12 | **46.32** | 44.08 | 41.47 | **57.76** | 54.18 | 48.90 | 22.95 | 43.84 | 43.34 | 42.17 | 54.96 | 41.97 |
| 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** |
| CNN [18] | 3 | 44.64 | 64.49 | 32.43 | 23.53 | 35.64 | 55.92 | 56.90 | 39.38 | 28.07 | 49.64 | 42.18 | 41.38 | 59.95 | 55.52 | 53.92 | 24.55 | 46.81 | 38.89 | 47.53 | 59.39 | 45.04 |
| R-CNN $\text{pool}_5$ [17] | - | 51.8 | 60.2 | 36.4 | 27.8 | 23.2 | 52.8 | 60.6 | 49.2 | 18.3 | 47.8 | 44.3 | 40.8 | 56.6 | 58.7 | 42.4 | 23.4 | 46.1 | 36.7 | 51.3 | 55.7 | 44.2 |

Table 1: Detection average precision (%) on VOC 2007 test. Column C shows the number of aspect ratio components. Performance improvement due to incorporation of context and multi-scale reasoning (MSS) with HOG and CNN features are shown. For reference, two other baselines, of a three aspect ratio components single-scale model and region proposal-based approach, are included. Note that the results of [18] for one and three aspect ratio components are using the publicly available code.

| | C | P | aero | bike | bird | boat | botl | bus | car | cat | chair | cow | table | dog | horse | mbike | pers | plant | sheep | sofa | train | tv | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C-DPM [41] | 3 | 8 | 39.7 | 59.5 | **35.8** | 24.8 | 35.5 | 53.7 | 48.6 | **46.0** | **29.2** | **36.8** | **45.5** | **42.0** | 57.7 | 56.0 | 37.4 | **30.1** | 31.1 | **50.4** | **56.1** | 51.6 | **43.4** |
| Conv-DPM [51] | 3 | 9 | **48.9** | **67.3** | 25.3 | **25.1** | **35.7** | **58.3** | **60.1** | 35.3 | 22.7 | 36.4 | 37.1 | 26.9 | **64.9** | **62.0** | **47.0** | 24.1 | **37.5** | 40.2 | 54.1 | **57.0** | 43.3 |

Table 2: The table depicts detection average precision (%) on VOC 2007 test for other methods employing **part modeling and CNN features**. The results are included for completeness, and meant to be compared with the results in Table 1. Our proposed method does not perform any explicit part reasoning.

| | C | aero | bike | bird | boat | botl | bus | car | cat | chair | cow | table | dog | horse | mbike | pers | plant | sheep | sofa | train | tv | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CNN | 1 | 41.48 | 62.58 | 36.88 | 16.65 | 22.23 | 48.07 | 61.31 | 50.78 | 29.41 | 49.10 | 47.54 | 45.64 | 62.45 | 58.13 | 50.61 | 25.57 | 48.58 | 48.01 | 44.81 | **59.53** | 45.47 |
| CNN-MSS | 1 | 46.68 | 58.09 | 33.83 | 15.48 | **29.62** | 47.41 | 63.58 | 51.34 | **31.97** | 48.19 | **54.71** | **46.11** | **60.29** | **59.66** | **51.01** | 24.87 | 48.65 | **50.89** | 44.91 | 57.35 | 46.23 |
| RCNN $\text{pool}_5$ [17] | - | **58.2** | **63.3** | **37.9** | **27.6** | 26.1 | **54.1** | **66.9** | **51.4** | 26.7 | **55.5** | 43.4 | 43.1 | 57.7 | 59.0 | 45.8 | **28.1** | **50.8** | 40.6 | **53.1** | 56.4 | **47.3** |
| RCNN $\text{fc}_7$ [17] | - | 64.2 | 69.7 | 50.0 | 41.9 | 32.0 | 62.6 | 71.0 | 60.7 | 32.7 | 58.5 | 46.5 | 56.1 | 60.6 | 66.8 | 54.2 | 31.5 | 52.8 | 48.9 | 57.9 | 64.7 | 54.2 |
| RCNN $\text{fc}_7$ BB [17] | - | 68.1 | 72.8 | 56.8 | 43.0 | 36.8 | 66.3 | 74.2 | 67.6 | 34.4 | 63.5 | 54.5 | 61.2 | 69.1 | 68.6 | 58.7 | 33.4 | 62.9 | 51.1 | 62.5 | 64.8 | 58.5 |

Table 3: Results with fine-tuned features on VOC 2007 test. Our approach uses no region proposals (unlike RCNN), a single aspect ratio model, and only $\text{conv}_5$ feature maps.

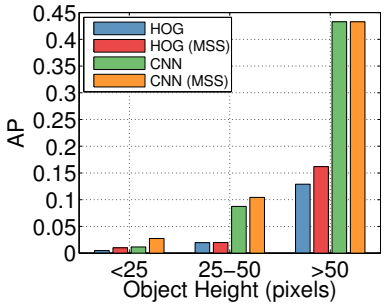



Figure 12: Improvement in performance for different object sizes. The largest gains due to incorporating the MSS approach are seen on smaller objects, which include more relevant contextual information throughout the multi-scale features.

**Comparison with state-of-the-art**: The main emphasis in this work is in analysis on modeling multi-scale context and its applications to efficient object detection and localization with deep features. The analysis framework was used to study scale importance, impact of dataset properties, and performance under varying object class and size settings. On PASCAL VOC, certain object classes greatly benefited from the proposed approach in detection, all of the 20 classes benefited in localization quality, and insights were made regarding challenging cases for the MSS approach. By employing only $\text{conv}_5$ feature maps, the method is efficient (requiring a single forward pass for each image scale) and have a low memory impact (no fully connected layers which contain most of the network parameters). As a reference, we provide absolute performance to other related research studies in Tables 1, 2, and 3 with different experimental settings.

For a fair comparison with a baseline, we closely followed Girshick *et al.* [18] in the deep feature pyramid extraction throughout the experiments. Overall, with a single aspect ratio model, our analysis results in a significant improvement of $4.48$ mAP over the results of [18], from $38.26$ mAP (obtained by the available implementation of [18]) to $42.74$. We observed model size to be a crucial parameter, and increasing it results in improvement of the baseline to $41.97$ mAP. Large gains in detection performance are shown for HOG, with an mAP increase of over 7 points. As discussed previously, the MSS approach has less

impact on objects with little scale variation. Furthermore, as multi-scale templates require scale-specific instances, a small number of instances in the dataset (even with data augmentation), leads to sub-optimal learning and reduced performance gains. On the other hand, certain classes (e.g. 'aeroplane', 'car', 'table', and 'sofa') show large gains in performance. As the method in [18] employs no contextual reasoning, a further gain is obtained by the multi-scale reasoning in overall mAP.

As a reference, although not the main focus of this study, the results of [18] with three aspect ratios are shown, which has an overall $6.78$ points improvement up to $45.04$ AP, improving over R-CNN in performance with the same convolutional feature maps. The improvement due to multiple aspect ratio components is an orthogonal improvement to MSS as context cues can be incorporated into each of the components. Furthermore, note that unlike R-CNN, [18] and our study does not involve a region proposal mechanism and per-region forward pass through the network (either through the whole network or just through the fully connected layers), which is computationally costly. The CNN-MSS approach ($42.74$ mAP) performs similarly to other recently proposed approaches of Wan *et al.* [51] and Savalle *et al.* [41] employing multiple aspect ratio components, CNN feature pyramids, and explicit part reasoning. The best relevant results is achieved with R-CNN, fine-tuning, multiple fully connected layers ($\text{fc}_7$), and bounding-box (BB) regression at $58.5$ mAP. Compared to R-CNN, the proposed approach is significantly more efficient in memory and computational cost. Furthermore, MSS learns scale-specific appearance and localization models while R-CNN does not. Results are shown both for no fine-tuning and with fine-tuning. R-CNN with the same convolutional features is outperformed on some classes where region proposals are weak. The results post fine-tuning shown in Table 3 demonstrate a consistent improvement. This is expected, as fine-tuning is mostly focused on improving local region representation.

**Run-time speed**: The computational speed is bound by two main factors, the feature pyramid extraction time and the model evaluation (either single-scale or MSS). The feature computation step (a 7 scale deep feature pyramid) is identical for the baseline and the MSS approach, running at $\sim 0.4$ seconds per image on PASCAL with a Titan X GPU. For the baseline, scoring a window $p_s$ using the features $\phi(p_s) \in \mathbb{R}^d$ involves $d$ operations, which is repeated over $S$ scales ($S \times d$). For a given image location, evaluation with the MSS detector involves $S$ models and an increase of the computational cost by a factor of $S$, to ($S \times S \times d$). In the current CPU implementation, the run-time of the MSS evaluation takes $\sim 0.7$ seconds per image. In the future, feature selection could potentially reduce the computational complexity of the detector evaluation for further speed gains.

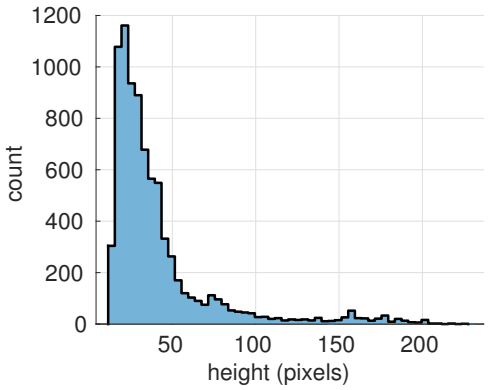
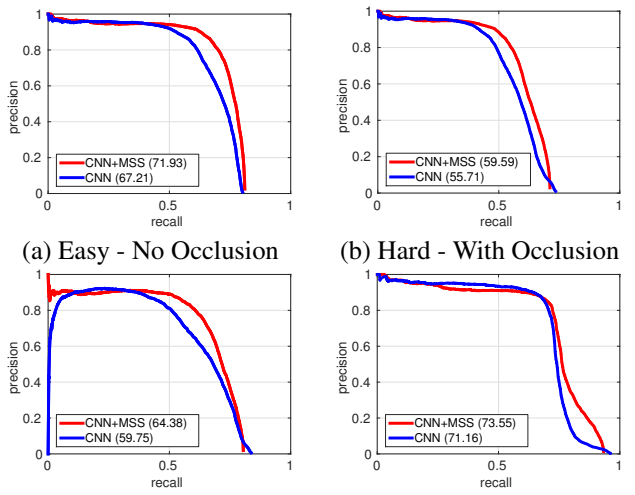



Figure 13: Height distribution in the highway vehicles dataset.

### 4.2. Results on Highway Vehicles

The PASCAL VOC 2007 dataset was used for developing the MSS approach and providing analysis in terms of impact of dataset properties, error types and localization quality, generalization to different object types, and sensitivity to object scale. In order to further test the performance of the proposed approach and understand its benefits, we employ a multi-view highway dataset captured using front and rear mounted cameras on a moving vehicle platform [9]. The highway settings are relevant as objects undergo large variation in scale as they enter and leave the scene. Furthermore, because the PASCAL VOC 2007 dataset targets generic object detection, it only contains a handful of images in settings similar to highway settings. The highway vehicles dataset is composed of a total of 1550 images containing 8295 objects. All truncated vehicles are also included in the evaluation. Object occlusion state have also been annotated in order to study performance under occlusion. The object height distribution is depicted in Fig. 13, showing large variation.

The results for vehicle detection are shown in Fig. 14. When occluded objects are excluded, the MSS approach results in a significant improvement of $4.72$ AP points over the baseline. With the inclusion of occluded objects, the improvement is consistent at $3.88$ AP points. On this dataset, a main improvement is in detecting smaller objects and better resolving multiple detection boxes, as shown in Fig. 14(c). By observing the curves in Fig. 14, we can see how the MSS approach maintains precision at a higher recall over the baseline. This is due to the improved multi-scale reasoning. While the baseline scores objects based on local information and therefore relies on the heuristic NMS alone to resolve responses at nearby locations and multiple scales, the MSS approach can better reason over responses in different scales. This can be clearly seen in the example images in Fig. 15, where detection results are shown for both




(a) Easy - No Occlusion (b) Hard - With Occlusion

(c) Height - 20-40 Pixels (d) Height - Above 40 Pixels

Figure 14: Results for vehicle detection on highway settings with different evaluation procedures.

the MSS and the single scale baseline at a fixed recall rate. Fig. 15 also shows cases where false positives are reduced due to contextual information available at multiple scales.

## 5. Concluding Remarks

In this paper, the role of multi-scale context in object detection with deep features was studied. An efficient framework for analysis of multi-scale contextual reasoning was proposed and studied on the PASCAL object detection benchmark and a highway vehicles dataset. Because the proposed approach operates on scale volumes, learns scale-specific models, and infers a localization label, it was shown to result in more robust detection and localization of objects. Visualization and feature selection analysis demonstrated how discriminative learning strongly favor multi-scale cues when these are present in training, both in adjacent and remote image scales. Comparative analysis evaluated generalization of the proposed approach for different feature types and dataset settings. As current state-of-the-art object detectors emphasize local region feature pooling in detection, the insights in this study can be used to train better CNN-based object detectors. In the future, the insights from this study will be used in order to design better end-to-end contextual, multi-scale detection frameworks.

## 6. Acknowledgments

We thank the reviewers and editors for their comments while preparing the manuscript, and thank our colleagues at the Computer Vision and Robotics Research Laboratory for their assistance. We also thank Zhuowen Tu for helpful discussions.

(a) CNN-MSS (proposed) (b) CNN (baseline)

Figure 15: Results for vehicle detection on highway settings at a fixed recall rate. Observe how the MSS approach better reasons over multi-scale responses, allowing for higher precision at the same recall rate and better localization compared to the single-scale CNN, which employs independent scoring at each scale and relies on NMS alone for resolving multi-scale responses.

## References

[1] 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.

[2] M. Blaschko and C. Lampert. Learning to localize objects

with structured output regression. In *European Conf. Computer Vision*, 2008.

[3] S. Branson, O. Beijbom, and S. Belongie. Efficient large-scale structured learning. In *IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional networks. In *British Machine Vision Conf.*, 2014.

[5] G. Chen, Y. Ding, J. Xiao, and T. X. Han. Detection evolution with multi-order contextual co-occurrence. In *IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[6] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *Intl. Journal Computer Vision*, 95(1):1–12, 2011.

[7] Y. Ding and J. Xiao. Contextual boost for pedestrian detection. In *IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

[8] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 36(8):1532–1545, 2014.

[9] J. V. Dueholm, M. S. Kristoffersen, R. K. Satzoda, E. Ohn-Bar, T. Moeslund, and M. M. Trivedi. Multi-perspective vehicle detection and tracking: Challenges, dataset, and metrics. In *IEEE Intl. Conf. Intelligent Transportation Systems*, 2016.

[10] 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.

[11] 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*, 88(2):303–338, 2010.

[12] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, 2013.

[13] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[14] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

[15] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[16] R. Girshick. Fast R-CNN. In *IEEE Intl. Conf. on Computer Vision*, 2015.

[17] 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.

[18] 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.

[19] 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.

[20] M. Hoai, L. Torresani, F. D. la Torre, and C. Rother. Learning discriminative localization from weakly labeled data. *Pattern Recognition*, 47(3):1523–1534, 2014.

[21] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *European Conf. Computer Vision*, 2012.

[22] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. In *IEEE Conf. Computer Vision and Pattern Recognition*. 2006.

[23] T. Joachims, T. Finley, and C.-N. Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59, 2009.

[24] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *Neural Information Processing Systems*, 2012.

[25] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *Intl. Conf. Machine Learning*, 2013.

[26] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[27] B. Li, T. Wu, and S.-C. Zhu. Integrating context and occlusion for car detection by hierarchical And-Or model. In *European Conf. Computer Vision*. 2014.

[28] C. Long, X. Wang, G. Hua, M. Yang, and Y. Lin. Accurate object detection with location relaxation and regionlets relocalization. In *Asian Conf. Computer Vision*, 2014.

[29] E. Ohn-Bar and M. M. Trivedi. Fast and robust object detection using visual subcategories. In *CVPRW*, 2014.

[30] E. Ohn-Bar and M. M. Trivedi. Learning to detect vehicles by clustering appearance patterns. *IEEE Trans. Intelligent Transportation Systems*, 16(5):2511–2521, 2015.

[31] D. Osaku, R. Nakamura, L. Pereira, R. Pisani, A. Levada, F. Cappabianco, A. Falco, and J. P. Papa. Improving land cover classification through contextual-based optimum-path forest. *Information Sciences*, 324:60–87, 2015.

[32] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. In *European Conf. Computer Vision*, 2010.

[33] B. Pepik, R. Benenson, T. Ritschel, and B. Schiele. What is holding back convnets for detection? In *German Conf. Pattern Recognition*, 2015.

[34] L. Quannan, J. Wang, Z. Tu, and D. P. Wipf. Fixed-point model for structured labeling. In *Intl. Conf. Machine Learning*, 2013.

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

[36] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Neural Information Processing Systems*, 2015.

[37] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5:3.

[38] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Technical report, DTIC Document, 1985.

[39] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *IEEE Conf. Computer Vision and Pattern Recognition*, 2011.

[40] M. A. Sadeghi and D. Forsyth. 30Hz object detection with DPM V5. In *ECCV*, 2014.

[41] P.-A. Savalle, S. Tsogkas, G. Papandreou, and I. Kokkinos. Deformable part models with cnn features. In *ECCVW*, 2014.

[42] 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.

[43] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

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

[45] B. Shi, X. Bai, and C. Yao. Script identification in the wild via discriminative convolutional neural network. *Pattern Recognition*, 52:448–458, 2016.

[46] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Intl. Conf. Learning Representations*, 2015.

[47] C. Szegedy, A. Toshev, and D. Erhan. Deep neural networks for object detection. In *Neural Information Processing Systems*, 2013.

[48] Z. Tu and X. Bai. Auto-context and its application to high-level vision tasks and 3D brain image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(10), 2010.

[49] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *Intl. Journal Computer Vision*, 104(2):154–171, 2013.

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

[51] L. Wan, D. Eigen, and R. Fergus. End-to-end integration of a convolutional network, deformable parts model and non-maximum suppression. In *IEEE Conf. Computer Vision and Pattern Recognition*, 2015.

[52] S. Xie and Z. Tu. Holistically-nested edge detection. In *IEEE Intl. Conf. on Computer Vision*, 2015.

[53] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Convolutional channel features. In *IEEE Intl. Conf. on Computer Vision*, 2015.

[54] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conf. Computer Vision*, 2014.

[55] N. Zhang, R. Farrell, F. Iandola, and T. Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *IEEE Intl. Conf. on Computer Vision*, 2013.

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

[57] Z. Zuo, G. Wang, B. Shuai, L. Zhao, and Q. Yang. Exemplar based deep discriminative and shareable feature learning for scene image classification. *Pattern Recognition*, 48(10):3004–3015, 2015.