

Learning a Layout Transfer Network for Context Aware Object Detection

Tao Wang^{ID}, Xuming He, Yuanzheng Cai, and Guobao Xiao^{ID}

Abstract—We present a context aware object detection method based on a retrieve-and-transform scene layout model. Given an input image, our approach first retrieves a coarse scene layout from a codebook of typical layout templates. In order to handle large layout variations, we use a variant of the spatial transformer network to transform and refine the retrieved layout, resulting in a set of interpretable and semantically meaningful feature maps of object locations and scales. The above steps are implemented as a Layout Transfer Network which we integrate into Faster RCNN to allow for joint reasoning of object detection and scene layout estimation. Extensive experiments on three public datasets verified that our approach provides consistent performance improvements to the state-of-the-art object detection baselines on a variety of challenging tasks in the traffic surveillance and the autonomous driving domains.

Index Terms—Object detection, context modeling, scene layout transfer.

I. INTRODUCTION

HUMAN perception almost always reflects an integration of information from bottom-up and top-down processes (e.g., [3]–[5]). In particular, context effects are an integral part of visual perception in humans. Consider the object detection problem as shown in Figure 1. As humans, we take for granted the ability to understand the layout of a scene at the first glance, and then know where to look for specific objects. For example, cars are most likely to appear on paved areas and those closer to the camera are typically larger than distant ones. It is therefore appealing to design models for object detection that reason about the spatial context in a similar fashion.

Manuscript received March 8, 2019; revised July 23, 2019; accepted August 28, 2019. Date of publication September 13, 2019; date of current version October 2, 2020. This work was supported in part by the NSFC under Grant 61703195 and Grant 61702431, in part by the Fujian NSF under Grant 2019J01756, in part by the Shanghai NSF under Grant 18ZR1425100, in part by The Education Department of Fujian Province under Grant JAT170459 and Grant JK2017039, in part by the Distinguished Young Scholars Program of Fujian Universities, in part by the Fuzhou Technology Planning Program under Grant 2018-G-96, and Grant 2018-G-98, and in part by the Minjiang University under Grant MJUKF201716, Grant MJY19021, and Grant MJY19022. The Associate Editor for this article was J. M. Alvarez. (*Corresponding author: Guobao Xiao.*)

T. Wang and Y. Cai are with the Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350108, China, also with the College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China, and also with NetDragon Inc., Fuzhou 350001, China.

X. He is with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China.

G. Xiao is with the Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350108, China (e-mail: gbx@mju.edu.cn).

Digital Object Identifier 10.1109/TITS.2019.2939213

On the contrary, recent state-of-the-art object detection algorithms produce scores for densely sampled object locations and scales, or a few hundreds to thousands of “blobby” object proposals. While it is true that these methods encode the spatial context to some extent due to the large receptive fields of the underlying convolutional neural nets (CNNs), methods like these usually lack a holistic understanding of the scene layouts. More importantly, unlike human perception, their performances deteriorate quickly when visual cues of objects become ambiguous or weak. Motivated by how humans understand the scene layouts for object detection, we propose a simple, interpretable and flexible framework for learning to transfer scene layouts for context-aware object detection.

The general idea of context modeling for object detection has long been proven effective in the computer vision community, with seminal works from Torralba [6], Murphy *et al.* [7], and Torralba *et al.* [8], and later Hoiem *et al.* [9], plus a few more [10]–[12] as prominent examples. More recently, the modeling of the spatial context has been extended to 3D scenarios [13]–[19] as high quality co-registered depth and color images have become more easily accessible. Most existing approaches assume a parametric model of the scene layout, such as the piecewise planar assumption [20], blocks world assumption [21], or the Manhattan world assumption [22]–[25], just to name a few. Despite the great progress, modeling the scene layout in a parametric fashion becomes challenging when the layout variation is large. In addition to an increased model complexity, performance could deteriorate quickly under atypical scene configurations. To address the above issues, we explore a semi-parametric approach to context modeling for object detection. In particular, we improve object detection through a coarse-to-fine scene layout model that predicts potential object locations and scales for every object category, as illustrated in Figure 1. Such a high dimensional mapping is difficult to learn in a purely parametric way due to the large output space, so we adopt a retrieve-and-transform strategy for scene layout prediction. Specifically, the input image (Figure 1 (a)) is first matched against a codebook of typical scene layout templates, and the most similar scene layout (Figure 1 (b) and (c)) is retrieved. This scene layout is further transformed and refined (Figure 1 (d) and (e)) to adapt to the specific appearance of the input image. In addition, we implement the above procedure as an integral component of Faster RCNN [2], so that the object detection and the scene layout estimation tasks can be learned in an alternating fashion.

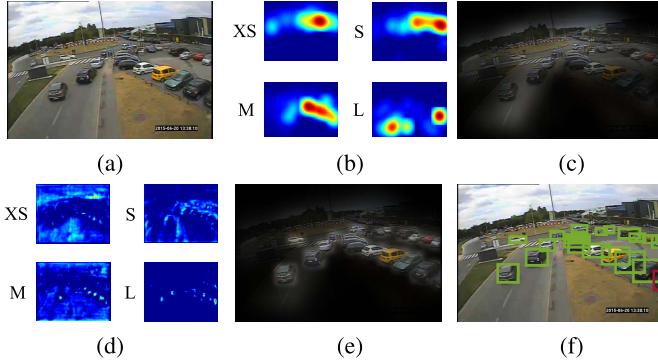


Fig. 1. Transferring scene layouts for object detection. (a) Input image. (b) Coarse scene layout for the *car* category at 4 different scales (XS, S, M, L) retrieved from similar images. (c) Object attention map derived from the coarse scene layout. (d) Scene layout after transformation and refinement. (e) Object attention map derived from the transformed and refined scene layout. (f) Detector output.

We note that the proposed method resembles human perception in the sense that we are able to seamlessly integrate information from bottom-up and top-down visual processings. The bottom-up processing involves the process that starts with deep image features; the top-down processing, which starts with retrieving scene layouts from our external scene layout memory, injects prior knowledge and expectations about the scene. By virtue of a simple and differentiable transformation, the proposed method is able to adapt the most relevant scene layout priors to a specific input image.

The benefit of our proposed method is threefold. Firstly, our scene layout representation is interpretable, which makes it more readily applicable to other scene understanding tasks. In fact, we show that we are able to build an external memory of typical scene layouts from a large database and then accurately retrieve the most relevant scene templates at test time. Secondly, a common problem in these retrieval-based methods is that it would be challenging to deal with intra-class variations. In our work, we use a variant of the spatial transformer network [1] to adapt scene layout templates to specific images, making our method capable of handling diverse scene layouts. Lastly, the proposed module can be integrated into a deep network for object detection, resulting in a single CNN for joint object detection and scene layout estimation. Extensive experiments on three public datasets verified that images in the traffic surveillance and automated driving domains are well-suited for our approach because their scene layouts provide strong priors for localizing objects.

This paper extends our previous work [26] in several ways. In particular, (i) scene layout estimation is now based on a drastically different method based on retrieve-and-transform, which is both more efficient and effective; (ii) scene layout estimation is now an integral part of the object detection CNN. This improves detection performance, and also removes unnecessary feature computational costs; (iii) we provide more detailed experimental evaluation and ablation studies on our proposed method in order to quantitatively motivate our model design. Furthermore, we compare our method against the baselines on two additional public datasets for object detection.

The rest of the paper is organized as follows. In Section II, we briefly discuss related work in object detection, context modeling, and scene layout estimation. The details of our model, including its structure, inference and learning, are introduced in Section III. This is followed by experimental evaluation in Section IV and closing remarks in Section V.

II. RELATED WORK

A. Object Detection

Recent years witnessed a huge success of Convolutional Neural Network (CNN) based object detection algorithms over conventional methods based on hand-crafted features and a shallow object grammar-based architecture. Some of the most prominent examples include sliding-window based OverFeat [27] and object proposal based R-CNN [28] and its faster variants [2], [29]–[31]. These methods are directly inspired by the success of CNN for image classification. The latter, proposal-based methods seek to exploit the strong representation power of deep networks to classify and make refinements to a relatively small set (typically hundreds to a few thousands) of potential object regions. Another line of work attempts to make direct predictions using a deep network without the object proposal step. Examples include YOLO [32], SSD [33], DSOD [34] and we note that these methods are generally more computationally efficient. In this work, we choose Faster RCNN [2] as our baseline object detector and explore how to improve their results via incorporating scene-level context cues.

B. Context Modeling

Context aware object detection has been well studied, and many context-aware object detection methods have been proposed (e.g., [6], [9]–[12], [35]–[38]). See [10] for a review and [39] for an empirical study of earlier work in the literature. More recently, Yang *et al.* [40] have shown that reasoning about a 2.1D layered object representation in a scene can positively impact object detection. Yao *et al.* [41] propose a holistic scene understanding model which jointly solves object detection, segmentation and scene classification. Mottaghi *et al.* [42] exploit both the local and global contexts by reasoning about the presence of contextual classes, and propose a context-aware improvement to the DPM [43]. Zhang *et al.* [44] propose a nonparametric column-based tiered model for scene layout estimation of road scenes. Zhu *et al.* [45] use CNNs to obtain contextual scores for object hypotheses, in addition to scores obtained with object appearance. Cai *et al.* [46] propose to ease the object scale variation issue by performing object detection with multiple CNN layers, each focusing on objects within certain scale ranges. Batzer *et al.* [47] propose a context-aware voting scheme for small and distant object detection. In addition, Sun and Jacobs [48] propose to learn a context model that predicts where objects may be missing. Other works have extended context modeling to 3D scenarios. For example, Bao, Sun and Savarase propose a parameterized 3D surface layout model and combine it with object detectors [14], [15]. Geiger *et al.* [49] propose a generative model for joint inference of scene topology, geometry and 3D object

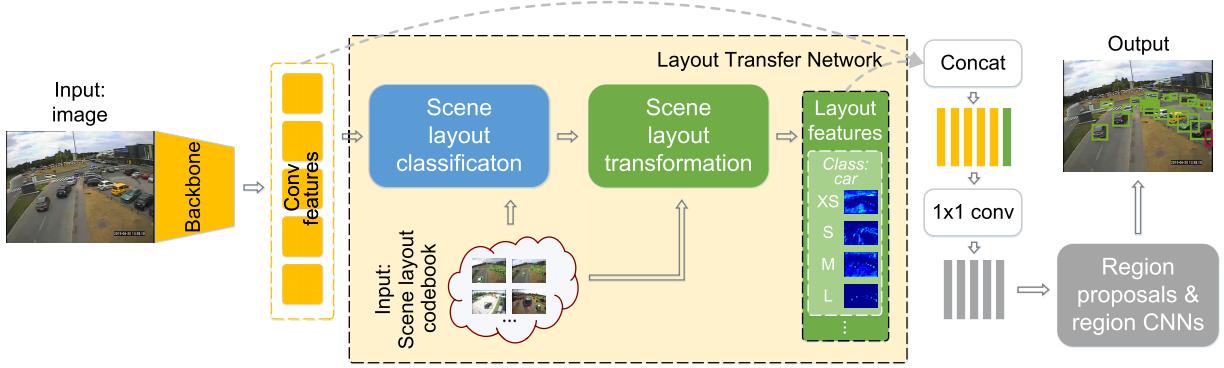


Fig. 2. The overall architecture of our method. Our layout transfer network adopts a retrieve-and-transform strategy for scene layout estimation. Given an input image, we first retrieve the most similar scene layout from a pre-built codebook with a classification sub-network (blue). The retrieved layout are then refined with a transformation sub-network, producing the final scene layout features (green). These features are then concatenated with backbone convolutional features (orange) to obtain the feature map for downstream object detection (gray).

locations. Wojek *et al.* [50] also propose a 3D scene model with explicit occlusion reasoning for object detection and tracking. Choi *et al.* [16] learn latent 3D geometric phrases to jointly solve object detection and scene layout estimation. Similarly, Lin *et al.* [17] use a CRF model to integrate various contextual relations for holistic scene understanding. Other later works include [18], [51] and [19]. Our work differs from the methods above in the sense that we propose a semi-parametric, retrieve-and-transform based approach to model the spatial context for object detection, which allows us to efficiently search within a high dimensional output space of the scene context. Our method is simple, interpretable, and it can be integrated into a deep network for object detection.

C. Scene Layout Estimation

Our work is also related to scene layout estimation methods that attempt to predict either parametric or nonparametric scene layout representations. For indoor scenes, recent works (e.g., [52]–[56]) made great progress by leveraging strong scene priors such as floor plans, geometric priors, or the more classical Manhattan world assumption, in conjunction with deep models trained on large-scale datasets. For outdoor scenes, Seff and Xiao [57] propose to use deep models to predict a set of driving-related road layout attributes. In addition, Mátyus *et al.* [58] propose a CRF-based scene layout model that combines perspective and top-view images. Zhai *et al.* [59] learn a transform to transfer the semantics from ground to the aerial image domain. Li *et al.* [60] propose a perspective-aware scene parsing method that estimates the perspective geometry of a scene image through a CNN to allow for finer parsing of small distant objects, and fuse prediction results at multiple scales with a perspective-aware CRF. Schulter *et al.* [61] propose a CNN that can hallucinate depth and semantics occluded by foreground objects and estimate a scene layout in the top view from a single perspective view. Wang *et al.* [62] propose a rich parametric top-view representation of complex road scenes that uses CNNs for predicting scene model parameters and a CRF for consistency reasoning. Contrary to the above methods, we propose a

simple and interpretable scene layout representation that can be directly used to improve object detection performance, and does not require additional data or annotation during training. Furthermore, we implement our retrieve-and-transform scene layout estimation model as part of an object detector that allows jointly learning for object detection and scene layout estimation.

III. OUR APPROACH

Let us begin with the definition of the object detection problem and notations. Given an input image, object detection algorithms output a score for each valid object hypothesis. More formally, suppose we have an input image $I \in \mathcal{I}$ where \mathcal{I} is the input image space. Let the object hypothesis be $\mathbf{x} \in \mathcal{X}$, where \mathcal{X} is the object hypothesis space. To simplify the notation, we assume each hypothesis is $\mathbf{x} = (\mathbf{x}_c, a_s, a_r, o)$ where $\mathbf{x}_c = (a_x, a_y)$ is the image coordinate location of the object center, a_s the scale, a_r the aspect ratio, and $o \in \mathcal{O}$ the object class. Note that each \mathbf{x} now implies a bounding box as well. Object detection algorithms define a scoring function $S(\mathbf{x}|I)$ for each valid object hypothesis \mathbf{x} . For example, this score is implemented as a two-class softmax score in Faster RCNN, i.e., $S(\mathbf{x}|I) = p(\mathbf{x}|I)$.

The core idea of this work is that we are able to transfer scene layouts from training images that are similar to the input image to predict potential object locations and scales. To this end, we propose an additional scene layout score $S_l(\mathbf{x}|\mathcal{N}_l)$ for any given object hypothesis \mathbf{x} by investigating a local neighborhood \mathcal{N}_l of the input image I . Directly learning such a high-dimensional mapping is difficult, so we adopt a two-step retrieve-and-transform strategy here:

- We first retrieve a coarse scene layout score $S_c(\mathbf{x}|\mathcal{C})$ by matching the input image I to a codebook \mathcal{C} of scene layout templates.
- Afterwards, we apply a transformation \mathcal{T} to $S_c(\cdot)$ to obtain the refined scene layout score, i.e., $S_l(\mathbf{x}|\mathcal{N}_l) = \mathcal{T}(S_c(\mathbf{x}|\mathcal{C}))$.

More specifically, denote the training dataset as $\mathcal{D} = \{(I_m, \mathcal{Y}_m, z_m)\}_{m=1}^M$ with M images, where I_m is a training

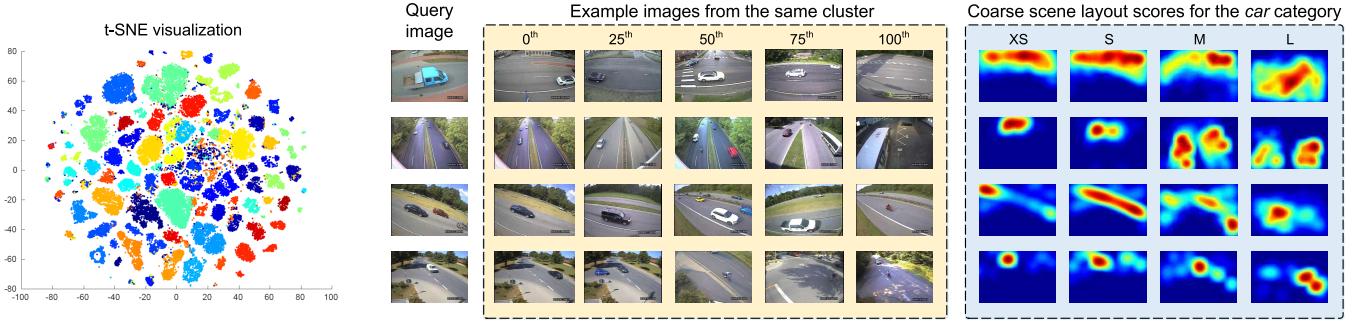


Fig. 3. **Left:** The t-SNE visualization [63] of the *pool5* features from the ResNet-50 [64] network, color-coded according to their cluster membership. **Right:** Given a query image (in the leftmost column), we present example images from the matching scene layout cluster on the MIO-TCD dataset (the yellow panel). In particular, we show images near the 0th, 25th, 50th, 75th and 100th percentiles according to their distances from the cluster center. In addition, we show coarse scene layout scores $S_c(\cdot)$ for the *car* category obtained on the MIO-TCD dataset (the blue panel). By matching a query image to our scene layout codebook, spatial distribution of object locations at 4 different scales (XS, S, M, L) are obtained. Note that the scene layout scores are still coarse, but they roughly suggest a scene layout in terms of potential object locations.

TABLE I
SUMMARY OF THE MAIN NOTATIONS. SEE SECTION III FOR DETAILS

Object detection	
input image	I
object hypothesis	$\mathbf{x} = (\mathbf{x}_c, a_s, a_r, o)$
object detection score	$S(\mathbf{x} I)$
Scene layout	
training dataset	$\mathcal{D} = \{(I_m, \mathcal{Y}_m, z_m)\}_{m=1}^M$
ground-truth annotation	$\mathbf{y} = (\mathbf{y}_c, a_s, a_r, o), \mathbf{y} \in \mathcal{Y}_m$
scene layout type	$z_m \in \mathcal{Z}, \mathcal{Z} = \{1, \dots, N\}$
image neighborhood	$\mathcal{N}_I, \mathcal{N}_I \in \mathcal{D}$
scene layout codebook	$\mathcal{C} = \{(z_m, \mathcal{Y}_m)\}_{m=1}^M$
coarse scene layout score	$S_c(\mathbf{x} \mathcal{C})$
refined scene layout score	$S_l(\mathbf{x} \mathcal{N}_I) = \mathcal{T}(S_c(\mathbf{x} \mathcal{C}))$

image and \mathcal{Y}_m is the set of all ground-truth annotations of the m -th image. Similar to \mathbf{x} , let $\mathbf{y} = (\mathbf{y}_c, a_s, a_r, o), \mathbf{y} \in \mathcal{Y}_m$ be a ground-truth object annotation. Instead of codewords, our scene layout codebook \mathcal{C} stores the scene layout *type* label z_m and the ground-truth annotations of each training image, i.e., $\mathcal{C} = \{(z_m, \mathcal{Y}_m)\}_{m=1}^M$, $z_m \in \mathcal{Z}$, where $\mathcal{Z} = \{1, \dots, N\}$ and N is the number of scene layout templates. In this way, each training image I_m is additionally associated with its scene layout type label z_m . The main notations are summarized in Table I.

There are three key issues here: (1) how to define the neighborhood \mathcal{N}_I , (2) how to obtain an appropriate representation for $S_l(\mathbf{x}|\mathcal{N}_I)$ that can be helpful for object detection, and (3) how to integrate $S_l(\mathbf{x}|\mathcal{N}_I)$ into an existing object detection framework. To address the first issue, we build a codebook of typical scene layouts by clustering the image-level appearance features. By matching an input image against the codebook, we can learn to transfer the neighborhood information encoded in the codebook entries to the target image. See Figure 3 for examples. For the second issue, we choose the object location heatmaps, since our ultimate goal is to predict object locations. In this work, we follow a retrieve-and-transform strategy to obtain a feature map that not only encodes the spatial distribution of objects in the image neighborhood, but also further adapted to specific images based on the input image appearance. The overall process is illustrated in Figure 2, with examples of the output scene layout scores shown

in Figure 7. For the last issue, we propose two strategies for information fusion either at the feature level or at the final decision scores, respectively.

It should be noted that our method provides interpretable intermediate features of the coarse layout $S_c(\cdot)$ and the refined layout $S_l(\cdot)$ that resemble the psychological process of humans looking up the memory of typical scene layouts, and then adapt them to specific scene appearances. This also allows us to inject additional supervision signals to help convergence during training. In addition, such a high-dimensional mapping is difficult to learn in a purely parametric way due to the large output space. We quantitatively show in the experiments that our strategy is much more effective than a naive baseline which attempts to directly predict the spatial locations of objects.

A. Building a Scene Layout Codebook

As the first foundation stone, we obtain scene layout templates in the training dataset by clustering the image-level appearance features. We build a codebook \mathcal{C} that encodes typical scene layouts, and train a classifier for the scene layout classification. At test time, we then classify the input image into one of these scene layout clusters, and obtain a rough estimate possible of object locations. To this end, we introduce a feature manifold that is descriptive of the scene layouts. Following [26], we use the 2048-D features extracted from the *pool5* layer of a ResNet-50 [64] network applied on the input image I , as it has been widely used in image classification to describe the appearance of an image. The network is pretrained on the ImageNet dataset [64], [65]. One advantage of these features is that they are pretrained on a large database and are potentially a more robust representation of the image appearance when compared to other alternatives.

More specifically, let M be the number of images in our training set and D be the feature dimension for the image-level appearance features (i.e., $D = 2048$ for our ResNet *pool5* features). We can perform K-means clustering with N clusters for the training feature matrix $\mathbf{F} \in \mathbb{R}^{D \times M}$. Denote z_m as the cluster membership for the m -th image in the training set, and $\delta(z_m = i)$ an indicator function for whether or not the m -th image belongs to the i -th cluster. Our scene layout

codebook stores the cluster membership labels and the ground-truth object annotations in a nonparametric fashion, i.e., $\mathcal{C} = \{(z_m, \mathcal{Y}_m)\}_{m=1}^M$. The scene layout score for the i -th cluster $S_c^i(\mathbf{x}|\mathcal{C})$ is obtained by accumulating all ground-truth object annotations in that cluster:

$$S_c^i(\mathbf{x}|\mathcal{C}) = \sum_{m=1}^M \frac{\delta(z_m = i)}{Z_c \sigma^2} \sum_{\mathbf{y} \in \mathcal{Y}_m} \exp\left(-\frac{\|\mathbf{y}_c - \mathbf{x}_c\|^2}{2\sigma^2}\right) \quad (1)$$

where $\mathbf{y} = (\mathbf{y}_c, a_s, a_r, o)$, $\mathbf{y} \in \mathcal{Y}_m$ is a ground-truth object annotation of the m -th image, and \mathcal{Y}_m is the set of all ground-truth object annotations in the m -th image. We build a mixture model with K components by sorting \mathbf{y} 's by their scale a_s , aspect ratio a_r , object class o and split them into K groups Y_1, \dots, Y_K . Each of these groups includes annotations within certain ranges of a_s and a_r , and of a specific object class o . Similar to \mathbf{y} , an object hypothesis \mathbf{x} can be sorted into one of these groups. For notation simplicity, we use $\mathbf{x} \simeq \mathbf{y}$ to denote that \mathbf{x} and \mathbf{y} are in the same group. This is straightforward because we would only like to accumulate support for an object hypothesis from those ground-truth annotations that are in the same group. In addition, \mathbf{y}_c and \mathbf{x}_c are the object center coordinates, and Z_c is a normalizing factor. Equation 1 additively accumulates support for an object hypothesis \mathbf{x} by looking at nearby spatial locations in the ground-truth object annotations while applying a Gaussian smoothing kernel.

The left part of Figure 3 presents the t-SNE visualization [63] of the training feature matrix \mathbf{F} , with each feature color-coded according to its cluster membership. The right part of Figure 3 shows a query image and example images (in the yellow panel) from the matching scene layout cluster on the MIO-TCD dataset. In particular, we show images near the 0th, 25th, 50th, 75th and 100th percentiles according to their distances from the cluster center. In this way, we are able to show the variations within individual clusters. Specifically, a scene layout cluster contains images taken with different cameras from similar views, not merely different images taken from the same camera. In addition, we show the spatial distribution of objects in the *car* category at 4 different scales (in the blue panel). We note that the heatmaps in Figure 3 can be viewed as a sampling distribution of object locations obtained from a specific image cluster, and they do not generally provide exact object locations for a given image. The intra-cluster layout variations are still large, and that we further refine the scene layout scores to adapt to the appearance of an input image before using them for object detection. We describe the relevant details in Section III-C. Before that, let us discuss details on how to retrieve the scores from the most similar scene layout cluster given a query image.

B. Scene Layout Classification

Now we move on to describe the first essential step towards scene layout estimation: scene layout classification. After building the scene layout codebook with N clusters, we are able to use the cluster membership as a class label for a training image. Mathematically, let \mathcal{I} be the input image space and $\mathcal{Z} = \{1, \dots, N\}$ be the label space. Let z_m be the

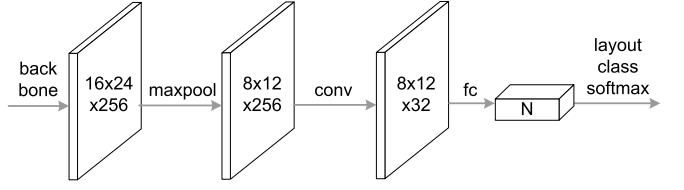


Fig. 4. The scene layout classification head. We use feature map dimensions on the MIO-TCD dataset as an example. A max pooling layer and a convolution layer are first applied to the feature backbone. Afterwards, a fully connected layer with N -way outputs predicts the scene layout cluster probabilities. N is the number of clusters in the scene layout codebook.

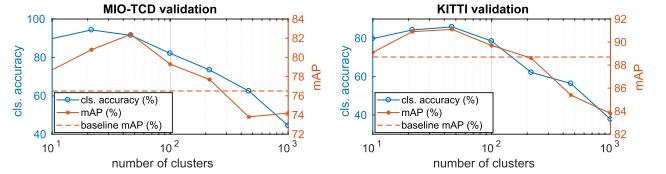


Fig. 5. The impact of number of clusters N on the scene layout classification accuracies and the mean APs. The left panel presents results on the MIO-TCD validation set, and the right panel presents results on the KITTI validation set.

cluster membership (also, the scene layout class label) for the m -th image in the training set. Therefore, we are given a classification dataset $\{(I_m, z_m)\}_{m=1}^M$ where each $(I_m, z_m) \in (\mathcal{I} \times \mathcal{Z})$. Our scene layout classifier maps the input image space to the label space $f_{cls} : \mathcal{I} \rightarrow \mathbb{R}^N$. We do so by reusing the backbone network of Faster RCNN, and adding an additional sub-network with a softmax output layer. The sharing of feature computation allows us to avoid unnecessary computational overheads. Importantly, we note that the scene layout classification allows us to retrieve the most relevant scene layouts in our codebook, thus making the downstream scene layout transformation and refinement easier to learn.

In terms of the network architecture, we attach a scene layout classification head to the backbone network, as shown in Figure 4. For simplicity in illustrations, we assume that the input image dimension is 512×768 , as we experimented with the MIO-TCD [66] dataset. Denote the output of the last residual blocks of $conv2$, $conv3$, $conv4$ and $conv5$ as $\{C_2, C_3, C_4, C_5\}$. Note that, due to the fully convolutional nature of the backbone, these feature maps have strides of $\{4, 8, 16, 32\}$ pixels with respect to the input image. For instance, the spatial dimensions of C_5 would be $16 \times 24 \times 256$ for a 512×768 input image, where 256 is the number of channels. We use C_5 as the input to our scene layout classification head. In order to obtain a more concise representation, our classification head starts with a 2×2 max pooling layer and a 3×3 stride 1 convolution layer, followed by a fully connected layer that integrates spatial information for classification.

We empirically found that setting the number of clusters N properly is important for obtaining the desired scene layout classification results, as well as object detection performance improvements. Figure 5 presents its impact on scene layout classification accuracies and mean APs for object detection on MIO-TCD and KITTI [67] validation sets, respectively. As we can see, if N is too small, the mean AP improvements

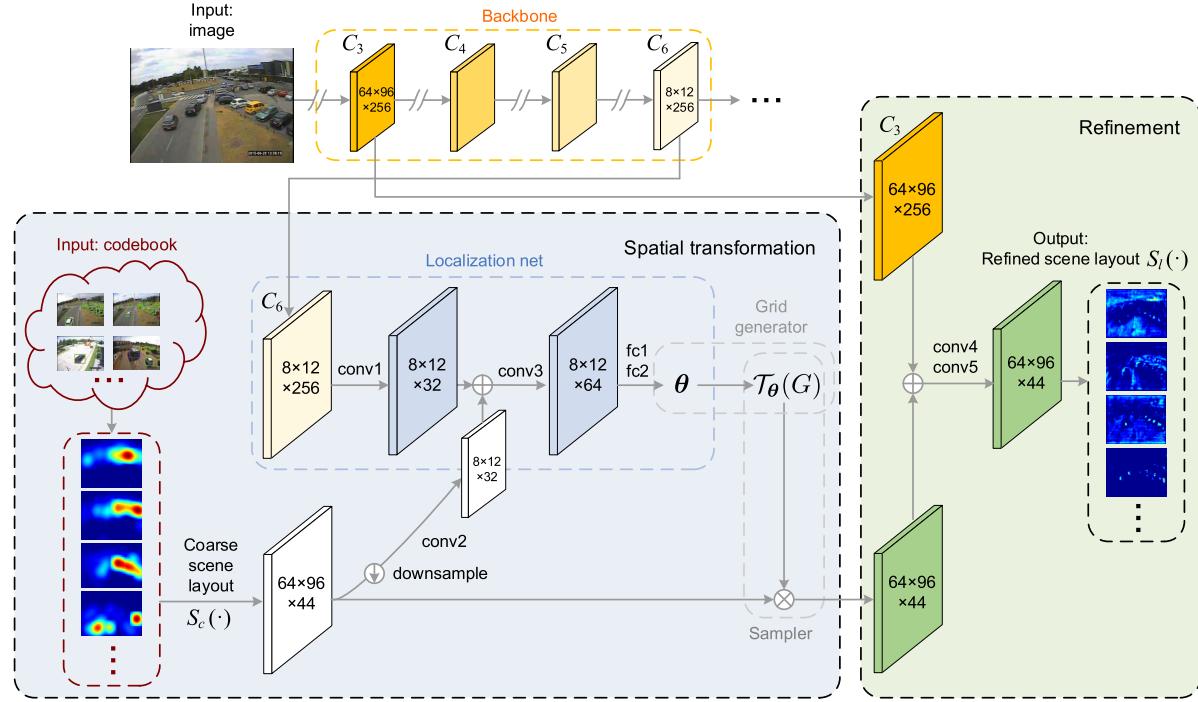


Fig. 6. The scene layout transformation sub-network. We use the feature map dimensions on the MIO-TCD dataset as an example. The sub-network consists of two modules: spatial transformation (the blue panel) and refinement (the green panel). Layer parameters are summarized in Table II.

may not be maximized and the classification is difficult due to the fact that very different scene layouts may be clustered into a same scene layout class. On the other hand, setting N to a large value can negatively impact both the classification accuracy and the mean AP. In general, we found that N is closely related to the number of typical scene layouts in a dataset. We set $N = 50$ for our experiments in this paper.

C. Scene Layout Transformation

As discussed in the previous section, we are able to retrieve a scene layout template in terms of object location heatmaps by matching an input image against the scene layout codebook. See Figure 3 for a few examples. Due to large intra-cluster scene layout variations, however, the retrieved templates are usually too coarse if directly used as a feature map for object detection. See Section IV-A for a detailed quantitative analysis. In order to handle the large layout variations, we propose a transformation and a refinement sub-network to adapt the scene layout template to the appearance of a specific input image. In particular, the spatial transformer network (STN) [1] proposes a generic CNN module that allows us to learn any feature transformation in a parameterized form, provided that it is differentiable with respect to the parameters.

Specifically, for an input image I suppose that the i -th cluster is the scene layout classification result (i.e., the i -th element in the output of the scene layout classification head has the largest logit), so that we retrieve the corresponding scene layout score $S_c^i(\mathbf{x}|\mathcal{C})$. We propose to use a variant of the STN to apply a transformation \mathcal{T} to $S_c(\cdot)$ to obtain the final scene layout score $S_l(\cdot)$, i.e., $S_l(\mathbf{x}|\mathcal{N}_I) = \mathcal{T}(S_c^i(\mathbf{x}|\mathcal{C}))$.

More specifically, \mathcal{T} consists of two consecutive CNN modules (namely, *spatial transformation* and *refinement*, as shown in Figure 6):

1) *Spatial Transformation*: Given the coarse scene layout, we first apply a 2D affine transformation to allow for rotation, translation, scale changes, etc. Because we use a mixture of discretized groups of scales and aspect ratios in this work, we can reduce the pose of an object hypothesis \mathbf{x} to its object center coordinates $\mathbf{x}_c = (a_x, a_y)$. Let (a_x, a_y) and (a'_x, a'_y) be the object center coordinates before and after the spatial transformation, respectively. In this case, the pointwise transformation in the homogeneous coordinates can be written as:

$$\begin{pmatrix} a_x \\ a_y \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} a'_x \\ a'_y \\ 1 \end{pmatrix} \quad (2)$$

As we can see from Equation 2, the 2D affine transformation \mathcal{T}_{θ} has 6 parameters. The STN estimates these parameters with a localization network $f_{loc} : (\mathcal{I} \times \mathcal{C}) \rightarrow \mathbb{R}^6$. In addition, we note that the transformed coordinates (a'_x, a'_y) are defined on a regular grid $G = \{(a'_x, a'_y)\}$ on an output feature map. Therefore, we can use a sampler to obtain their corresponding coordinates $\mathcal{T}_{\theta}(G)$ in (a_x, a_y) by applying Equation 2.

Figure 6 presents the detailed network architecture of our spatial transformation module. For simplicity, the feature dimensions on the MIO-TCD dataset are shown as an example. The dataset has 11 object categories, and in this paper we use 4 scale groups and a single aspect ratio group for each object category, so the number of channels for $S_c(\cdot)$ and $S_l(\cdot)$ is $11 \times 4 \times 1 = 44$. At the top of Figure 6 is the feature backbone,

which is shared with downstream object detection modules. As before, we denote the output of the last residual blocks of *conv2*, *conv3*, *conv4* and *conv5* as $\{C_2, C_3, C_4, C_5\}$. Additionally, denote C_6 as the feature map that is downsampled by half from C_5 with a 2×2 max pooling layer. There are two important backbone feature maps we are using in our scene layout transformation sub-network. The first one is C_6 , whose spatial dimension is $8 \times 12 \times 256$ and has a stride of 64 pixels with respect to the input image. This layer can be used as a feature that is indicative of the overall appearance of images, and with a relatively small number of parameters. The other one is C_3 , whose spatial dimension is $64 \times 96 \times 256$ and has a stride of 8 pixels with respect to the input image. We will use it as a lower level (i.e., finer) appearance feature that helps our model to attend to specific object locations in the refinement module. The spatial resolution of C_3 is identical to those object location heatmaps $S_c(\cdot)$ we store with the codebook \mathcal{C} in our implementation. We note that, however, this is only a particular implementation choice; there may be other potentially appropriate choices.

The localization network f_{loc} is shown in the blue dashed box in Figure 6. In order to regress the localization parameters θ , we use both the coarse scene layout $S_c(\cdot)$ and the backbone features C_6 . The idea is that the transformation parameters will be learned from both the retrieved (coarse) scene layout, and the appearance of the input image. $S_c(\cdot)$ is first downsampled from $64 \times 96 \times 44$ to $8 \times 12 \times 44$. The resultant feature maps and C_6 are then bottlenecked to 32 channels (with *conv1* and *conv2*, respectively) and then concatenated before going through a convolutional layer (*conv3*) and two fully connected layers (*fc1* and *fc2*) in order to regress θ . θ is then supplied to the grid generator and the sampler, which are standard components in an STN [1].

2) *Refinement*: In addition to the affine transformation, we would like the refined scene layout $S_l(\cdot)$ to be able to have an attention mechanism that focuses on specific regions within a particular input image that are likely to contain objects (see Figure 1). As this would require lower level image features for more accurate localization, we make use of C_3 , and concatenate it with the affine-transformed scene layout before going through two additional convolutional layers (*conv4* and *conv5*), to obtain the final transformed scene layout $S_l(\cdot)$. We note that *conv4* and *conv5* are different to feature maps C_4 and C_5 in Figure 6 as they learn dedicated feature transformations in order to obtain the refined scene layout $S_l(\cdot)$. On the contrary, C_4 and C_5 are the shared backbone features. The module is shown in the green dashed box in Figure 6. The specific layer parameters in our scene layout transformation sub-network are summarized in Table II.

Table III presents some performance comparisons among using different backbone features for our scene layout transformation sub-network. For the spatial transformation module, we experimented with C_6 , C_4 and $C_6 + C_4$, respectively. In order to use C_4 for our localization network, we downsample the feature map so that it has the same spatial resolution to C_6 , before going through the downstream processings as shown in Figure 6. We left out C_5 because C_6 is directly obtained from C_5 through max pooling. The use of C_4 allows

TABLE II
LAYER PARAMETERS OF THE SCENE LAYOUT TRANSFORMATION
SUB-NETWORK USED ON THE MIO-TCD DATASET. SEE
FIGURE 6 FOR THE NETWORK ARCHITECTURE

	layer parameters	output size
Spatial transformation		
<i>conv1</i>	$3 \times 3 \times 32$, stride 1	$8 \times 12 \times 32$
<i>conv2</i>	$3 \times 3 \times 32$, stride 1	$8 \times 12 \times 32$
<i>conv3</i>	$3 \times 3 \times 64$, stride 1	$8 \times 12 \times 64$
<i>fc1</i>	$6144 \rightarrow 512$	512
<i>fc2</i>	$512 \rightarrow 6$	6
Refinement		
<i>conv4</i>	$3 \times 3 \times 256$, stride 1	$64 \times 96 \times 256$
<i>conv5</i>	$3 \times 3 \times 44$, stride 1	$64 \times 96 \times 44$

TABLE III
AVERAGE PRECISION VALUES WE OBTAINED ON THE MIO-TCD
VALIDATION SET WHILE USING VARIOUS COMBINATIONS
OF BACKBONE FEATURES IN OUR LAYOUT
TRANSFORMATION SUB-NETWORK

backbone-features	AP	AP ₅₀	AP ₇₅
Spatial transformation			
C_6	57.5	80.9	64.8
C_4	52.2	78.2	57.9
$C_6 + C_4$	55.9	80.0	62.8
Refinement			
C_3	57.5	80.9	64.8
$C_3 + C_4$	57.2	80.8	64.2
$C_3 + C_4 + C_5$	56.8	80.5	63.9

us to assess the efficacy of lower level features when it comes to estimating the spatial transformation parameters. It is clear that C_6 performs better than the alternatives. For the refinement module, we tested with C_3 , $C_3 + C_4$ and $C_3 + C_4 + C_5$, respectively. C_4 and C_5 are upsampled where necessary. The idea is to verify if a cascade of feature maps at multiple feature pyramid levels would additionally benefit the object detection performance. We empirically found that using C_3 alone is sufficient to provide a strong performance. We note that, however, in this case C_4 and C_5 are still used for object detection as backbone features. In general, although the exact connectivity pattern may vary beyond the cases being discussed above, we observe consistent performance improvements when our layout transformation sub-network is used. See Section IV for details.

We present some examples from the outputs of our scene layout transformation in Figure 7, which shows the potential locations for the *car* category at 4 different scales.

D. Piecing Things Together

The scene layout predictions, as shown in Figure 7, provide useful context cues for object detection. In this paper, we propose two strategies to integrate these cues into an object detection framework:

1) *Late Fusion*: Following [26], we can directly use the scene layout scores in conjunction with the output of an object detector. The final object detection score is a weighted sum of

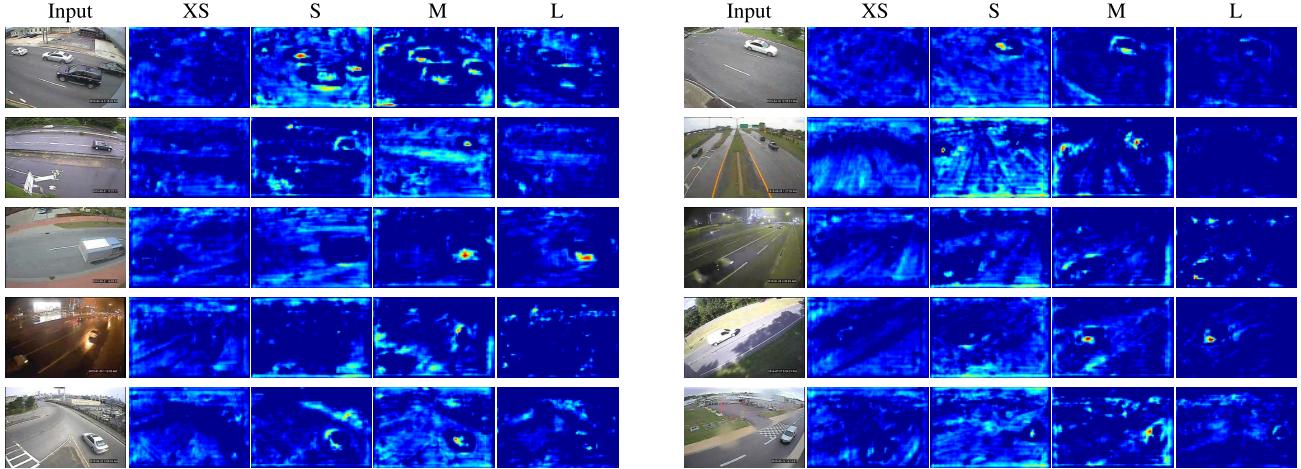


Fig. 7. Examples of the refined scene layout scores $S_l(\cdot)$ for the *car* category obtained with the MIO-TCD dataset. Input images are shown in the leftmost column, with locations for extra small (XS, farthest), small (S, far), medium (M, close) and large (L, closest) objects in the four columns to the right.

the two scores:

$$S(\mathbf{x}) = S_d(\mathbf{x}|I) + \alpha S_l(\mathbf{x}|\mathcal{N}_I) \quad (3)$$

where $S_d(\mathbf{x}|I)$ is the scoring function of an object detector, such as Faster RCNN. α is a hyperparameter for the relative importance between the two terms.

2) *Early Fusion*: We use the scene layout scores as an intermediate representation that enhances our image features. In particular, we experimented with various feature fusion methods, as shown in Table VIb. Our best-performing model uses an 1×1 conv layer, as illustrated in Figure 2. Specifically, for each level in the feature pyramid [68], we resize the scene layout scores to the corresponding feature map resolution with bilinear interpolation, and then concatenate the scores to the feature map. The 1×1 convolution then maps the features back to its original dimensions before concatenation. The main advantage of the early fusion strategy is that it results in a model that allows for alternating optimization of object detection and scene layout estimation. In Section IV, we compare early fusion with late fusion on the three datasets evaluated in this paper. The details of the training schedule for early fusion are presented in Section III-F.

E. Parameter Learning

In this section, we discuss details pertaining to the learning of the scene layout transfer for object detection. Particularly, the interpretable nature of our scene layout classification and transformation sub-networks allows us to inject supervision signals during training. We do so by adding additional terms to the learning objective of the object detection algorithm. We begin by discussing our overall learning objective.

1) *Learning Objective*: Denote the loss function of our baseline object detection algorithm as \mathcal{L}_{det} . In the case of Faster RCNN [2], this is a multi-task learning objective that involves object classification and bounding box regression. The overall learning objective of our proposed method can be written as follows:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{stn}} \quad (4)$$

where \mathcal{L}_{cls} and \mathcal{L}_{stn} are the loss functions for scene layout classification and transformation, respectively.

2) *Learning Scene Layout Classification*: For scene layout classification, we use the multi-class cross entropy loss, which is the most widely used loss function for neural network based classification. Specifically, the scene layout classification loss \mathcal{L}_{cls} is defined as:

$$\mathcal{L}_{\text{cls}} = - \sum_{m=1}^M \sum_{i=1}^N \delta(z_m = i) \log f_{\text{cls}}^i(I_m) \quad (5)$$

where $\delta(\cdot)$ is the indicator function and $f_{\text{cls}}^i(I_m)$ denotes the i -th element of $f_{\text{cls}}(I_m)$.

3) *Learning Scene Layout Transformation*: The output of our scene layout transformation sub-network, $S_l(\mathbf{x}|\mathcal{N}_I)$, is a set of heatmaps for object locations of different object categories at various scales and aspect ratios. We assume that the spatial dimensions of $S_l(\cdot)$ are $W \times H \times K$, where W and H are the width and height of the heatmaps, and $K = N_o \times N_s \times N_{ar}$ is the number of components in the mixture model discussed in Section III-A, e.g., $K = 11 \times 4 \times 1 = 44$ in the case of the MIO-TCD dataset. Here N_o , N_s and N_{ar} are the number of object categories, scale groups, and aspect ratio groups, respectively. Although it is possible to train the scene layout transformation sub-network without additional supervision, we are able to learn the network parameters with target scene layouts derived from the ground-truth annotations, which we denote as $S_l^*(\cdot)$. Overall, we write our scene layout transformation loss as follows:

$$\mathcal{L}_{\text{stn}} = \mathcal{L}_{\text{lyt}} + \beta \mathcal{L}_{\text{reg}} \quad (6)$$

where the layout loss, \mathcal{L}_{lyt} , quantifies the mismatch between $S_l(\cdot)$ and $S_l^*(\cdot)$, and \mathcal{L}_{reg} is a regularization term to account for the fact that the affine transformation shall not drift too far from an identity mapping. Here β is a hyperparameter for the tradeoff between the two terms.

The layout loss \mathcal{L}_{lyt} is an element-wise mean squared error (MSE) loss given by:

$$\mathcal{L}_{\text{lyt}} = \frac{1}{WHK} \sum_{m=1}^M \sum_{a_x=1}^W \sum_{a_y=1}^H \sum_{k=1}^K (S_{l,m}^*(\mathbf{x}|\mathcal{N}_{I_m})_{a_x,a_y,k} - S_{l,m}(\mathbf{x}|\mathcal{N}_{I_m})_{a_x,a_y,k})^2 \quad (7)$$

TABLE IV
FEATURE MAP RESOLUTIONS AND THE KERNEL DIMENSIONS
OF $conv3$ IN THE TRANSFORMATION SUB-NETWORK.
SEE SECTION III-F FOR DETAILS

dataset	input	C_3	C_6	$conv3$
MIO-TCD	512×768	64×96	8×12	$3 \times 3 \times 64$
Traffic lights	736×1280	92×160	12×20	$3 \times 3 \times 32$
KITTI	512×1664	64×208	8×26	$3 \times 3 \times 32$

Here the target scene layout output of the m -th image, $S_{l,m}^*(\mathbf{x}|\mathcal{N}_{I_m})$, is obtained by accumulating ground-truth object annotations in that image:

$$S_{l,m}^*(\mathbf{x}|\mathcal{N}_{I_m}) = \frac{1}{Z_l \sigma^2} \sum_{\substack{\mathbf{x} \simeq \mathbf{y} \\ \mathbf{y} \in \mathcal{Y}_m}} \exp \left(-\frac{\|\mathbf{y}_c - \mathbf{x}_c\|^2}{2\sigma^2} \right) \quad (8)$$

where $\mathbf{y} = (\mathbf{y}_c, a_s, a_r, o)$, $\mathbf{y} \in \mathcal{Y}_m$ is a ground-truth object annotation, and \mathcal{Y}_m is the set of all ground-truth object annotations in the m -th image. $\mathbf{x} \simeq \mathbf{y}$ denotes that \mathbf{x} and \mathbf{y} are in the same mixture model component. Here \mathbf{x}_c and \mathbf{y}_c are the object center coordinates, and Z_l is a normalizing factor. Finally, the regularization term \mathcal{L}_{reg} is given by:

$$\mathcal{L}_{\text{reg}} = \sum_{m=1}^M \frac{1}{N_\theta} \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_m\|_2^2 \quad (9)$$

where $\boldsymbol{\theta}_m = [\theta_{11} \ \theta_{12} \ \theta_{13} \ \theta_{21} \ \theta_{22} \ \theta_{23}]^T_m$ are the parameters of the affine transformation and $\boldsymbol{\theta}^* = [1 \ 0 \ 0 \ 0 \ 1 \ 0]^T$ is the identity transformation. $N_\theta = 6$ is the number of elements in $\boldsymbol{\theta}$.

F. Training Details

In this paper, we adopt a 3-step training schedule to learn object detection and scene layout estimation in an alternating fashion. In the first step, we train the Faster RCNN detector following the standard practice described in the paper [2]. A scene layout codebook is also learned offline. In the second step, we freeze the backbone features, and learn the scene layout classification sub-network and the scene layout transformation sub-network *sequentially*. This order is important because without an accurate scene layout being retrieved, the training of the transformation sub-network would become unstable. In this step, we use the SGD solver with initial learning rates set to 0.0025. In the third step, we fine-tune the object detector again with the scene layout classification and transformation sub-networks fixed. The initial learning rate is reduced to 0.001. The above alternating training can be run for more iterations, but we observed negligible improvements.

Due to the fully convolutional nature of the backbone network, the dimensions of the feature maps will depend on the resolution of the input image. The settings being used in this paper are summarized in Table IV. For the Bosch Small Traffic Lights [69] and the KITTI datasets, the dimensions of C_6 are higher than that of the MIO-TCD dataset, so we reduce the number of convolutional kernels by half in $conv3$ (i.e., $64 \rightarrow 32$), in order to reduce the number of parameters

in the $fc1$ layer of the transformation sub-network. We note that, however, that our model is somewhat robust to such changes in the network architecture. It should also be noted that, the higher input image resolution is essential for the Bosch Small Traffic Lights dataset, as most of the objects are smaller than 32×32 pixels. In addition, the start size of RPN anchors are also reduced from 32 to 8 to better cope with small objects.

IV. EXPERIMENTAL RESULTS

In this section, we thoroughly compare the proposed method with state-of-the-art object detection algorithms on public benchmark datasets, and present ablation studies to verify the efficacy of our model design choices. We focus on two important sub-areas in vision processing for intelligent transportation systems: traffic surveillance and autonomous driving. We evaluate the proposed method on three challenging object detection datasets:

- **MIOvision Traffic Camera Dataset (MIO-TCD)** [66]: To the best of our knowledge, MIO-TCD is the largest public benchmark for object detection in traffic surveillance images, with 110,000 images for training and 27,743 images for testing for its detection task. For this test set, the dataset offers a public object detection challenge: the TSWC-2017 localization challenge. It allows participating teams to upload their test results to an evaluation server to obtain their detection performance. Ranking of entries are based on AP₅₀, following the PASCAL VOC challenge [70].
- **Bosch Small Traffic Lights Dataset** [69]: The Bosch Small Traffic Lights dataset presents a unique challenge for detecting small objects with partial occlusions. The training and testing tests contain 5,093 and 8,334 images, respectively. Context information can be particularly useful for localizing small objects when the appearance cues are weak from objects themselves.
- **KITTI Object Detection** [67]: The KITTI object detection dataset has 7,481 training images and 7,518 testing images. As our method crucially relies on the availability of large-scale datasets that cover objects at various locations and scales, we evaluate the detection performance of the most prevalent *car* category that has 28,742 training instances. The KITTI dataset also offers a public online benchmark based on AP₇₀.

We refer our method to Layout Transfer Network (LTN) in the following sections.

A. Results on the MIO-TCD Dataset

In order to perform detailed ablation studies, we randomly split the original training set of the MIO-TCD dataset into a smaller training set with 99,000 images and a held-out validation set with 11,000 images. All results reported in this section are obtained by training models on the smaller training set, with ablation studies being carried out on the held-out validation set.

TABLE V

PER-CLASS AND MEAN AVERAGE PRECISION VALUES (AP_{50} , IN %) WE OBTAINED IN THE TSWC-2017 LOCALIZATION CHALLENGE (MIO-TCD TEST SET). OUR ENTRY OBTAINED A MEAN AP_{50} OF 82.75, WHICH IS THE STATE-OF-THE-ART. WE HAVE THE HIGHEST AP_{50} IN 8 OUT OF 11 OBJECT CATEGORIES AMONG ALL ENTRIES

MIO-TCD-test	backbone	a.truck	bicycle	bus	car	motorcycle	m.vehicle	n.m.vehicle	pedestrian	p.truck	s.u.truck	workvan	mean AP_{50}
Challenge submissions													
LTN (Ours)	R101-FPN	92.38	88.42	97.64	95.14	92.65	70.65	58.41	67.31	92.74	74.97	79.93	82.75
Jung et al. [71]	R101-ensemble	92.48	87.34	97.46	89.70	88.21	62.32	59.09	48.57	92.25	74.42	79.86	79.24
Wang et al. [26]	VGG-16+context	91.62	79.90	96.77	93.80	83.63	56.40	58.23	42.61	92.75	73.80	79.56	77.19
Baseline approaches													
SSD-512 [33]	VGG-16	91.28	77.36	96.56	93.59	79.53	55.39	56.60	41.58	92.66	72.74	79.40	76.06
YOLO v2 [72]	DarkNet-19	88.31	78.64	95.13	81.36	81.36	51.70	56.57	24.96	86.48	69.23	76.43	71.83
Faster RCNN [2]	VGG-16	80.70	70.63	93.45	79.85	74.58	46.48	21.22	19.49	86.71	53.29	67.40	63.07

TABLE VI

ABLATIONS. WE REPORT AVERAGE PRECISION VALUES FROM OUR ABLATION STUDIES ON THE MIO-TCD VALIDATION SET. THE BACKBONE NETWORK IS A RESNET-101-FPN [68]. SEE SECTION IV-A FOR DISCUSSIONS

early-vs-late-fusion	AP	AP ₅₀	AP ₇₅	fusion-method	AP	AP ₅₀	AP ₇₅	layout-resolution	AP	AP ₅₀	AP ₇₅
LTN – late fusion	55.2	79.5	61.9	eltwise-mul	56.2	80.0	63.3	32 × 48	55.4	79.8	61.9
LTN – early fusion	57.5	80.9	64.8	eltwise-sum	54.7	79.5	61.2	64 × 96	57.5	80.9	64.8
	+2.3	+1.4	+2.9	1 × 1 conv	57.5	80.9	64.8	128 × 192	57.7	80.9	65.1

(a) **Early vs. late fusion:** We compare two fusion methods for the learned scene layout scores. Early fusion allows for joint training for object detection and scene layout estimation, and compares favorably to late fusion.

(b) **Fusion method:** For early feature-level fusion, we experimented with different network structures. 1×1 conv outperforms other alternatives.

(c) **Layout resolution:** AP results with various layout resolutions. Resolutions above 64×96 (i.e., the resolution of the C_3 feature map) give diminishing returns.

	classify?	transform?	refine?	AP	AP ₅₀	AP ₇₅
Baseline	✗	✗	✗	48.2	76.1	52.7
	✓	✗	✗	52.0	77.6	58.0
+ components	✓	✓	✗	53.3	78.8	59.6
	✓	✗	✓	55.5	79.8	62.2
Full model	✓	✓	✓	57.5	80.9	64.8

(d) **Component test:** To isolate the performance gains, we switch off the following three components from our full model: (1) scene layout classification, (2) spatial transformation, and (3) refinement. Each component provides its own AP improvements, while the full model performs the best.

(e) **Fully convolutional network (FCN) vs. Layout transfer network (LTN):** Directly predicting scene layouts with an FCN does not produce superior results.

(f) **Layout regularization:** The layout regularization term is essential for learning the spatial transformation. We observe large performance gaps without the regularizer being used.

The quantitative results we obtained in the TSWC-2017 localization challenge¹ (i.e., the MIO-TCD test set) are reported in Table V. Comparing to other entries, our method shows a clear advantage. In particular, our results are 3.51 points better in terms of AP_{50} than the current winning entry from Jung *et al.* [71]. It should be noted, however, that their method uses a 4-models ensemble (two ResNet-50s and two ResNet-101s) based on R-FCN [31] and ours use only a single model. In addition, our results are 5.56 AP_{50} points better than our previous work [26] that uses a nonparametric label transfer method to predict the scene layout for object detection. Overall, we obtain the highest average precisions on 8 out of 11 object categories. To the best of our knowledge, these

results are the state-of-the-art in the MIO-TCD localization challenge.

In addition to the results on the test set, we also report some ablations we obtained on the held-out validation set, which are presented in Table VI. Specifically, we perform the following ablation studies:

- **Early vs. late fusion** (Table VIa): We compare the performance of early and late fusion as discussed in Section III-D. As expected, the early fusion strategy allows us to jointly optimize for object detection and scene layout estimation, resulting in better detection performance.
- **Fusion method** (Table VIb): In the case of early fusion, there are various ways to integrate the inferred scene layout features $S_l(\cdot)$ with backbone image features. In this

¹<http://podoce.dinf.usherbrooke.ca/results/localization>

set of experiments, we explore different ways to feature fusion, including elementwise multiplication (eltwise-mul), elementwise summation (eltwise-sum), and 1×1 convolution. The 1×1 convolution variant performs best and is used in all other experiments by default. We note that, for the two elementwise fusion methods, we need to apply an 1×1 convolution layer before the elementwise operations to make sure that the feature dimensions of the backbone and the scene layout are compatible.

- **Layout resolution** (Table VIc): Because we use a pyramid of features in the FPN backbone, a higher scene layout resolution could be helpful for detecting and localizing smaller objects. In general, a layout resolution at 64×96 for the MIO-TCD dataset (i.e., the resolution of C_3) is sufficient and higher resolutions give marginal performance gains. In this work, we stick to the resolution of C_3 as the scene layout resolution on all three datasets, as it provides a good tradeoff between performance and model complexity (i.e., the number of parameters).
- **Component test** (Table VID): The three main modules in the scene layout estimation are: (1) scene layout classification, (2) spatial transformation, and (3) refinement. Compared to a baseline Faster RCNN that uses none of these components, adding each module or any combination of them contribute to the AP performance.
- **FCN vs. LTN** (Table VIE): One of the main advantages of our proposed method is that we are able to retrieve a rough estimate of the scene layout in the first step, making the subsequent transformation and refinement easier to learn. Here we also report results obtained by learning a fully convolutional network that directly predicts the scene layout features $S(\cdot)$ from the backbone features C_3 and C_6 . As we can see, our retrieve-and-transform strategy boosts the detection performance by a reasonable margin.
- **Layout regularization** (Table VII): Finally, we verify the effectiveness of our layout regularization term in Equation 9. We observe a large performance improvement with the regularizer being used. In addition, we found in our experiments that the regularization term helps stabilize training and avoid over-fitting.

It is aforementioned that our method provides an interpretable representation of the scene layouts. In Figure 7, we present some examples of the scene layouts for the *car* category at 4 different scales as predicted by our method. It is clear from these examples that our scene layout scores can reliably predict object locations and scales through the object location heatmaps. In addition, we present some qualitative detection results in Figure 8. As we can see, our method is able to reliably detect overlapping and distant objects by incorporating the context cues. Typical failure modes are also presented in the last two rows, among the most common are class confusions, false alarms, and missing out-of-context objects.

B. Results on the Bosch Small Traffic Lights Dataset

In addition, we report the results we obtained on the Bosch Small Traffic Lights dataset. This dataset presents a unique

TABLE VII
AVERAGE PRECISION VALUES WE OBTAINED ON THE
BOSCH SMALL TRAFFIC LIGHTS TEST SET

<i>Traffic-lights-test</i>	AP	AP ₅₀	AP ₇₅
Faster RCNN	30.25	73.44	15.91
Faster RCNN + LTN (late fusion)	31.20	75.83	18.52
Faster RCNN + LTN (early fusion)	31.64	75.97	19.06
+ LTN (late fusion)	+1.0	+2.4	+2.6
+ LTN (early fusion)	+1.4	+2.5	+3.2

TABLE VIII
AVERAGE PRECISION VALUES (AP₇₀, IN %) WE OBTAINED
ON THE KITTI VALIDATION SET

<i>KITTI-val</i>	Moderate	Easy	Hard
Faster RCNN	88.72	90.76	80.31
Faster RCNN + LTN (late fusion)	93.12	94.04	86.25
Faster RCNN + LTN (early fusion)	93.09	94.42	86.65
+ LTN (late fusion)	+4.4	+3.3	+5.9
+ LTN (early fusion)	+4.4	+3.7	+6.3

challenge to detecting small objects with partial occlusions in autonomous driving. The detection results we obtained on the test set are reported in Table VII. In particular, accurate localization for objects as small as a few pixels wide (the medium width is about 8.6 pixels) is very challenging for state-of-the-art object detectors, as reflected by the low AP₇₅ values in a sharp contrast to the AP₅₀ values. Although we found that temporal information may be used to improve accuracy on this dataset (as some traffic lights may be missed while being detected in preceding or succeeding frames), it is not the focus of this paper and is therefore not used. Once again, our layout transfer network is able to improve detection results in these very challenging scenarios. Also, we emphasize again that using high resolution input images, as specified in Table IV, and smaller RPN anchors are vital to obtain satisfactory results on this particular dataset. We present some example detection results in Figure 9.

C. Results on the KITTI Dataset

The KITTI dataset provides a comprehensive set of real-world and challenging computer vision tasks including stereo, optical flow, visual odometry, object detection and tracking for scene understanding in autonomous driving. A 2D object detection dataset with 7,481 annotated training images is provided. As it is vital for our method to transfer scene layouts from a large dataset with objects at various locations and scales, we only evaluate the detection performance of *cars*, which is the largest object category. Other object categories (i.e., *pedestrians* and *cyclists*) are much smaller in comparison, so we do not include them in our evaluation. In addition, we split the original training set into a separate training set and a validation set with 5,985 and 1,496 images, respectively. This allows us to directly demonstrate the effectiveness of LTN, as compared to a baseline with identical settings otherwise.

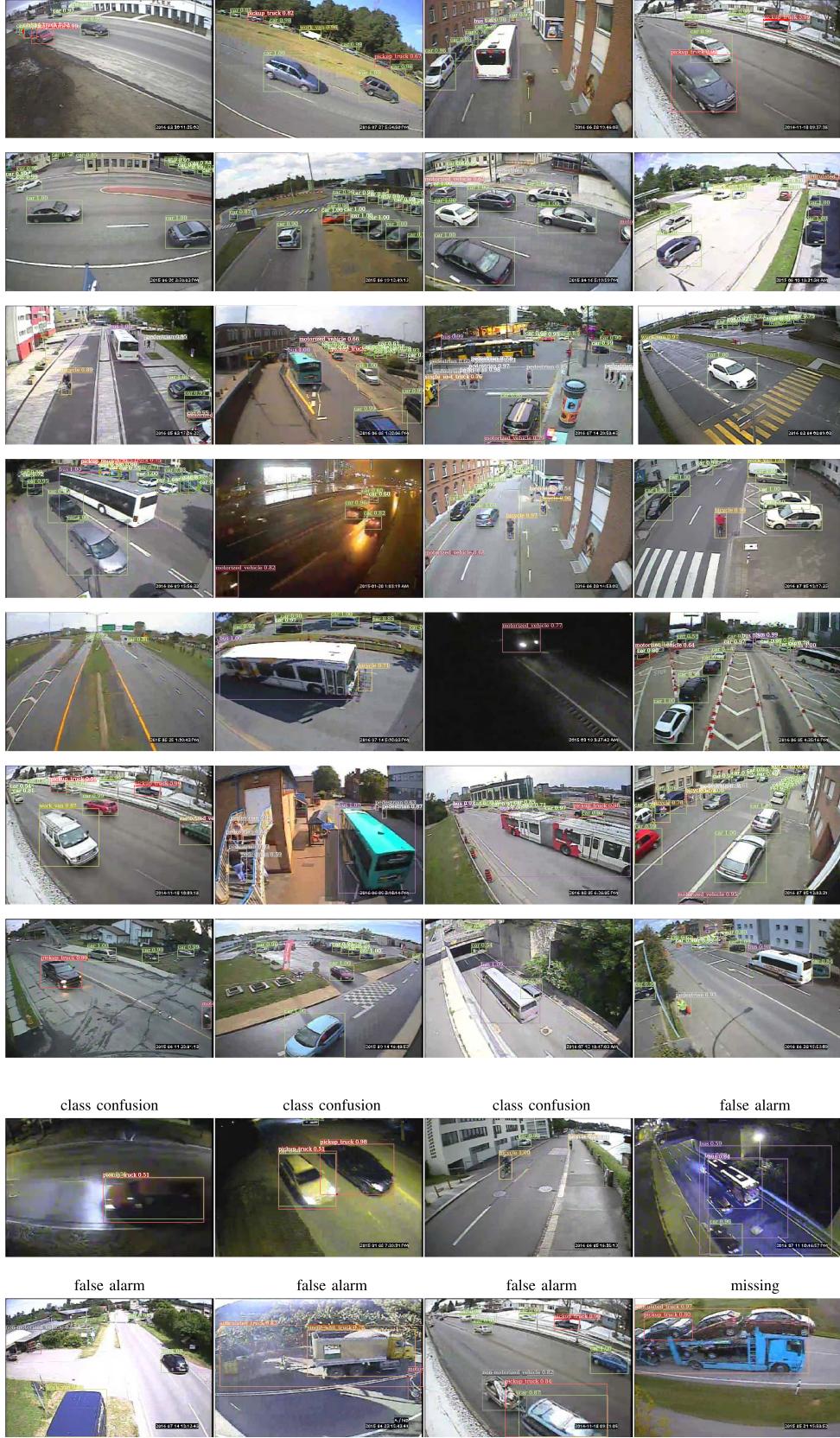


Fig. 8. Example detection results on the MIO-TCD test set. The last two rows present some typical failure modes, namely class confusion, false alarm, and missing out-of-context objects. See Section IV-A for details. Best viewed electronically, zoomed in.

Following the evaluation protocol of the dataset, the AP₇₀ values are reported on *Moderate*, *Easy*, and *Hard* objects, respectively. Table VIII summarizes the results on the KITTI

validation set. Compared to the baseline, our LTN (with early fusion) improves the average precision by 4.4, 3.7 and 6.3 points on the three difficulty levels, respectively.



Fig. 9. Example detection results on the Bosch Small Traffic Lights test set. See Section IV-B for details. Best viewed electronically, zoomed in.



Fig. 10. Example detection results on the KITTI test set. See Section IV-C for details. Best viewed electronically, zoomed in.

TABLE IX

AVERAGE PRECISION VALUES (AP₇₀, IN %) ON THE KITTI BENCHMARK. METHODS ARE RANKED BASED ON THEIR PERFORMANCE AT THE MODERATE DIFFICULTY LEVEL. IN GENERAL, THE LTN ACHIEVES A GOOD TRADEOFF BETWEEN ACCURACY AND SPEED, ESPECIALLY AT THE HARD DIFFICULTY LEVEL.
SEE TEXT FOR DETAILS

KITTI-test	time/image	AP ₇₀		
		Moderate	Easy	Hard
RRC [73]	3.6s	90.23	90.61	87.44
SJTU-HW [78], [79]	0.85s	90.08	90.81	79.98
SINet [74]	0.2s	89.60	90.60	77.75
Deep3DBox [80]	1.5s	89.04	92.98	77.17
MS-CNN [46]	0.4s	89.02	90.03	76.11
Mono3D [81]	4.2s	88.66	92.33	78.96
SDP+CRC (ft) [82]	0.6s	83.53	90.33	71.13
spLBP [83]	1.5s	77.40	87.19	60.60
Reinspect [84]	2s	76.65	88.13	66.23
Regionlets [85], [86], [87]	1s	76.45	84.75	59.70
SubCat [88]	0.7s	75.46	84.14	59.71
LTN (Ours)	0.4s	88.85	90.12	79.62

In particular, the performance on the *Hard* difficulty level is boosted by a large margin, as contextual cues are important for detecting objects that are heavily occluded or truncated

by image boundaries. In addition, Table IX reports the results from the KITTI benchmark (i.e., test set). In general, our method achieves a good tradeoff between accuracy and speed, especially at the *Hard* difficulty level. In particular, the top-performing method RRC [73] is around 10 times slower than our method due to the recurrent nature of its architecture. Another state-of-the-art method, SINet [74], is faster than our approach but we perform better at the *Hard* difficulty level. In addition, we note that the contributions of others may be orthogonal to ours. Furthermore, some state-of-the-art methods on the online benchmark ² are not listed here because they either require additional input modalities (e.g., PC-CNN-V2 [75] and F-PointNet [76]) or CAD models during training (e.g., Deep MANTA [77]).

We present some qualitative results in Figure 10. As we can see, scene layouts can be well indicative of potential vehicle locations and scales, and that our method is able to detect overlapping and distant objects with high accuracy.

V. CONCLUSION

In this paper, we proposed a layout transfer network for context aware object detection. An important aspect of our method is that we are able to obtain an interpretable scene

²<http://www.cvlibs.net/datasets/kitti>

layout representation which can be directly used to improve object detection performance. The scene layout transfer in our method provides a general approach to context modeling for object detection that can be used in conjunction with many other detection algorithms not mentioned in this paper. In the future, we wish to achieve scene layout classification with an integrated deep learning approach. In particular, this may allow for integrated training of all sub-systems and parameters, which may provide better overall performance. For example, we can use backpropagation to fine-tune the scene layout codebook after the initialization by clustering. We hope that our work would serve as a modest spur to induce further exploration into simple and robust scene layout representations that may be useful for a wider variety of scene understanding problems.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their insightful comments. They also thank Zhiming Luo and Pierre-Marc Jodoin for their help in the authors' participation in the TSWC-2017 challenge.

REFERENCES

- [1] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. NIPS*, 2015, pp. 2017–2025.
- [2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015, pp. 91–99.
- [3] J. L. McClelland and D. E. Rumelhart, "An interactive activation model of context effects in letter perception: I. An account of basic findings," *Psychol. Rev.*, vol. 88, no. 5, p. 375, 1981.
- [4] S. E. Palmer, *Vision Science: Photons to Phenomenology*. Cambridge, MA, USA: MIT Press, 1999.
- [5] T. S. Lee and D. Mumford, "Hierarchical Bayesian inference in the visual cortex," *J. Opt. Soc. Amer. A, Opt. Image Sci. Vis.*, vol. 20, no. 7, pp. 1434–1448, 2003.
- [6] A. Torralba, "Contextual priming for object detection," *Int. J. Comput. Vis.*, vol. 53, no. 2, pp. 169–191, 2003.
- [7] K. Murphy, A. Torralba, and W. T. Freeman, "Using the forest to see the trees: A graphicalmodel relating features, objects, and scenes," in *Proc. NIPS*, 2003, pp. 1–8.
- [8] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, "Context-based vision system for place and object recognition," in *Proc. ICCV*, Oct. 2003, pp. 273–280.
- [9] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," *Int. J. Comput. Vis.*, vol. 80, no. 1, pp. 3–15, 2008.
- [10] L. Wolf and S. Bileschi, "A critical view of context," *Int. J. Comput. Vis.*, vol. 69, no. 2, pp. 251–261, 2006.
- [11] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *Proc. ICCV*, Oct. 2007, pp. 1–8.
- [12] M. B. Blaschko and C. H. Lampert, "Object localization with global and local context kernels," in *Proc. BMVC*, 2009, p. 2.
- [13] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Depth from familiar objects: A hierarchical model for 3D scenes," in *Proc. CVPR*, Jun. 2006, pp. 2410–2417.
- [14] S. Y.-Z. Bao, M. Sun, and S. Savarese, "Toward coherent object detection and scene layout understanding," in *Proc. CVPR*, Jun. 2010, pp. 65–72.
- [15] M. Sun, Y. Bao, and S. Savarese, "Object detection with geometrical context feedback loop," in *Proc. BMVC*, 2010, p. 2.
- [16] W. Choi, Y.-W. Chao, C. Pantofaru, and S. Savarese, "Understanding indoor scenes using 3D geometric phrases," in *Proc. CVPR*, Jun. 2013, pp. 33–40.
- [17] D. Lin, S. Fidler, and R. Urtasun, "Holistic scene understanding for 3D object detection with RGBD cameras," in *Proc. ICCV*, Dec. 2013, pp. 1417–1424.
- [18] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proc. ECCV*, 2014, pp. 345–360.
- [19] W. Liu, R. Ji, and S. Li, "Towards 3D object detection with bimodal deep Boltzmann machines over RGBD imagery," in *Proc. CVPR*, 2015, pp. 3013–3021.
- [20] O. D. Faugeras and F. Lustman, "Motion and structure from motion in a piecewise planar environment," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 2, no. 3, pp. 485–508, 1988.
- [21] A. Gupta, A. A. Efros, and M. Hebert, "Blocks world revisited: Image understanding using qualitative geometry and mechanics," in *Proc. ECCV*, 2010, pp. 482–496.
- [22] V. Hedau, D. Hoiem, and D. Forsyth, "Recovering the spatial layout of cluttered rooms," in *Proc. ICCV*, Sep./Oct. 2009, pp. 1849–1856.
- [23] D. C. Lee, M. Hebert, and T. Kanade, "Geometric reasoning for single image structure recovery," in *Proc. CVPR*, Jun. 2009, pp. 2136–2143.
- [24] A. Gupta, M. Hebert, T. Kanade, and D. M. Blei, "Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces," in *Proc. NIPS*, 2010, pp. 1288–1296.
- [25] Y.-W. Chao, W. Choi, C. Pantofaru, and S. Savarese, "Layout estimation of highly cluttered indoor scenes using geometric and semantic cues," in *Proc. Int. Conf. Image Anal. Process.*, 2013, pp. 489–499.
- [26] T. Wang, X. He, S. Su, and Y. Guan, "Efficient scene layout aware object detection for traffic surveillance," in *Proc. CVPRW*, Jul. 2017, pp. 926–933.
- [27] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. ICLR*, 2014, pp. 1–16.
- [28] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. CVPR*, Jun. 2014, pp. 580–587.
- [29] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. ECCV*, 2014, pp. 346–361.
- [30] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, Dec. 2015, pp. 1440–1448.
- [31] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. NIPS*, 2016, pp. 379–387.
- [32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. CVPR*, Jun. 2016, pp. 779–788.
- [33] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. ECCV*, 2016, pp. 21–37.
- [34] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue, "DSOD: Learning deeply supervised object detectors from scratch," in *Proc. ICCV*, Oct. 2017, pp. 1919–1927.
- [35] A. Torralba, K. P. Murphy, and W. T. Freeman, "Contextual models for object detection using boosted random fields," in *Proc. NIPS*, 2004, pp. 1401–1408.
- [36] S. Kluckner, T. Mauthner, P. M. Roth, and H. Bischof, "Semantic image classification using consistent regions and individual context," in *Proc. BMVC*, 2009, p. 7.
- [37] M. Maire, S. X. Yu, and P. Perona, "Object detection and segmentation from joint embedding of parts and pixels," in *Proc. ICCV*, Nov. 2011, pp. 2142–2149.
- [38] J. Pan and T. Kanade, "Coherent object detection with 3D geometric context from a single image," in *Proc. ICCV*, Dec. 2013, pp. 2576–2583.
- [39] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *Proc. CVPR*, Jun. 2009, pp. 1271–1278.
- [40] Y. Yang, S. Hallman, D. Ramanan, and C. Fowlkes, "Layered object detection for multi-class segmentation," in *Proc. CVPR*, Jun. 2010, pp. 3113–3120.
- [41] J. Yao, S. Fidler, and R. Urtasun, "Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation," in *Proc. CVPR*, 2012, pp. 702–709.
- [42] R. Mottaghi *et al.*, "The role of context for object detection and semantic segmentation in the wild," in *Proc. CVPR*, Jun. 2014, pp. 891–898.
- [43] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
- [44] D. Zhang, X. He, and H. Li, "Data-driven street scene layout estimation for distant object detection," in *Proc. DICTA*, Nov. 2014, pp. 1–7.

- [45] Y. Zhu, R. Urtasun, R. Salakhutdinov, and S. Fidler, "segDeepM: Exploiting segmentation and context in deep neural networks for object detection," in *Proc. CVPR*, 2015, pp. 4703–4711.
- [46] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. ECCV*, 2016, pp. 354–370.
- [47] A.-K. Batzer, C. Scharfenberger, M. Karg, S. Lueke, and J. Adamy, "Generic hypothesis generation for small and distant objects," in *Proc. 19th IEEE Int. Conf. Intell. Transp. Syst.*, Nov. 2016, pp. 2171–2178.
- [48] J. Sun and D. W. Jacobs, "Seeing what is not there: Learning context to determine where objects are missing," in *Proc. CVPR*, Jul. 2017, pp. 5716–5724.
- [49] A. Geiger, C. Wojek, and R. Urtasun, "Joint 3D estimation of objects and scene layout," in *Proc. NIPS*, 2011, pp. 1467–1475.
- [50] C. Wojek, S. Walk, S. Roth, K. Schindler, and B. Schiele, "Monocular visual scene understanding: Understanding multi-object traffic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 882–897, Apr. 2013.
- [51] S. Wang, S. Fidler, and R. Urtasun, "Holistic 3D scene understanding from a single geo-tagged image," in *Proc. CVPR*, Jun. 2015, pp. 3964–3972.
- [52] A. Mallya and S. Lazebnik, "Learning informative edge maps for indoor scene layout prediction," in *Proc. ICCV*, Dec. 2015, pp. 936–944.
- [53] S. Dasgupta, K. Fang, K. Chen, and S. Savarese, "Delay: Robust spatial layout estimation for cluttered indoor scenes," in *Proc. CVPR*, Jun. 2016, pp. 616–624.
- [54] C. Liu, A. G. Schwing, K. Kundu, R. Urtasun, and S. Fidler, "Rent3D: Floor-plan priors for monocular layout estimation," in *Proc. CVPR*, Jun. 2015, pp. 3413–3421.
- [55] I. Armeni *et al.*, "3D semantic parsing of large-scale indoor spaces," in *Proc. CVPR*, Jun. 2016, pp. 1534–1543.
- [56] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proc. CVPR*, Jul. 2017, pp. 1746–1754.
- [57] A. Seff and J. Xiao, "Learning from maps: Visual common sense for autonomous driving," 2016, *arXiv:1611.08583*. [Online]. Available: <https://arxiv.org/abs/1611.08583>
- [58] G. Mátyus, S. Wang, S. Fidler, and R. Urtasun, "HD maps: Fine-grained road segmentation by parsing ground and aerial images," in *Proc. CVPR*, Jun. 2016, pp. 3611–3619.
- [59] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, "Predicting ground-level scene layout from aerial imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 867–875.
- [60] X. Li *et al.*, "FoveaNet: Perspective-aware urban scene parsing," in *Proc. ICCV*, Oct. 2017, pp. 784–792.
- [61] S. Schulter, M. Zhai, N. Jacobs, and M. Chandraker, "Learning to look around objects for top-view representations of outdoor scenes," in *Proc. ECCV*, 2018, pp. 815–831.
- [62] Z. Wang, B. Liu, S. Schulter, and M. Chandraker, "A parametric top-view representation of complex road scenes," in *Proc. CVPR*, Jun. 2019, pp. 10325–10333.
- [63] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
- [64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.
- [65] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
- [66] Z. Luo *et al.*, "MIO-TCD: A new benchmark dataset for vehicle classification and localization," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5129–5141, Oct. 2018.
- [67] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.
- [68] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. CVPR*, Jul. 2017, pp. 936–944.
- [69] K. Behrendt, L. Novak, and R. Botros, "A deep learning approach to traffic lights: Detection, tracking, and classification," in *Proc. ICRA*, May/Jun. 2017, pp. 1370–1377.
- [70] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*. Accessed: Sep. 6, 2019. [Online]. Available: <http://host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html>
- [71] H. Jung, M.-K. Choi, J. Jung, J.-H. Lee, S. Kwon, and W. Y. Jung, "ResNet-based vehicle classification and localization in traffic surveillance systems," in *Proc. CVPRW*, Jul. 2017, pp. 934–940.
- [72] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. CVPR*, 2017, pp. 7263–7271.
- [73] J. Ren *et al.*, "Accurate single stage detector using recurrent rolling convolution," in *Proc. CVPR*, Jul. 2017, pp. 5420–5428.
- [74] X. Hu *et al.*, "SINet: A scale-insensitive convolutional neural network for fast vehicle detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 1010–1019, Mar. 2019.
- [75] X. Du, M. H. Ang, S. Karaman, and D. Rus, "A general pipeline for 3D detection of vehicles," in *Proc. ICRA*, May 2018, pp. 3194–3200.
- [76] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3D object detection from RGB-D data," in *Proc. CVPR*, Jun. 2018, pp. 918–927.
- [77] F. Chabot, M. Chaouch, J. Rabarissoa, C. Teuliére, and T. Chateau, "Deep MANTA: A coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image," in *Proc. CVPR*, Jul. 2017, pp. 2040–2049.
- [78] S. Zhang, X. Zhao, L. Fang, H. Fei, and H. Song, "LED: Localization-quality estimation embedded detector," in *Proc. ICIP*, Oct. 2018, pp. 584–588.
- [79] L. Fang, X. Zhao, and S. Zhang, "Small-objectness sensitive detection based on shifted single shot detector," *Multimedia Tools Appl.*, vol. 78, pp. 13227–13245, May 2019.
- [80] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3D bounding box estimation using deep learning and geometry," in *Proc. CVPR*, Jul. 2017, pp. 7074–7082.
- [81] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3D object detection for autonomous driving," in *Proc. CVPR*, Jun. 2016, pp. 2147–2156.
- [82] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers," in *Proc. CVPR*, Jun. 2016, pp. 2129–2137.
- [83] Q. Hu, S. Paisitkriangkrai, C. Shen, A. van den Hengel, and F. Porikli, "Fast detection of multiple objects in traffic scenes with a common detection framework," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 1002–1014, Apr. 2016.
- [84] R. Stewart, M. Andriluka, and A. Y. Ng, "End-to-end people detection in crowded scenes," in *Proc. CVPR*, Jun. 2016, pp. 2325–2333.
- [85] C. Long, X. Wang, G. Hua, M. Yang, and Y. Lin, "Accurate object detection with location relaxation and regionlets re-localization," in *Proc. ACCV*, 2014, pp. 260–275.
- [86] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2071–2084, Oct. 2015.
- [87] W. Y. Zou, X. Wang, M. Sun, and Y. Lin, "Generic object detection with dense neural patterns and regionlets," in *Proc. BMVC*, 2014, pp. 1–11.
- [88] E. Ohn-Bar and M. M. Trivedi, "Learning to detect vehicles by clustering appearance patterns," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2511–2521, May 2015.



Tao Wang received the B.E. degree in information engineering from the South China University of Technology, Guangzhou, China, in 2009, and the Ph.D. degree in computer science from The Australian National University, Canberra, ACT, Australia, in 2016. He was a member of the Computer Vision Research Group, National ICT Australia, Canberra. He is currently a Lecturer with the College of Computer and Control Engineering, Minjiang University, Fuzhou, China. He is also with the College of Mathematics and Computer Science, Fuzhou University, Fuzhou, and NetDragon Inc., Fuzhou. His research interests include scene understanding, object detection, and semantic instance segmentation.



Xuming He received the Ph.D. degree in computer science from the University of Toronto, Toronto, ON, Canada, in 2008. He held a Post-Doctoral position with the University of California at Los Angeles, Los Angeles, CA, USA, from 2008 to 2010. He was an Adjunct Research Fellow with The Australian National University, Canberra, ACT, Australia, from 2010 to 2016. He joined National ICT Australia, Canberra, where he was a Senior Researcher from 2013 to 2016. He is currently an Associate Professor with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. His research interests include semantic image and video segmentation, 3-D scene understanding, visual motion analysis, and efficient inference and learning in structured models.



Guobao Xiao received the Ph.D. degree in computer science and technology from Xiamen University, China, in 2016. He was a Post-Doctoral Fellow with the School of Aerospace Engineering, Xiamen University, from 2016 to 2018. He is currently a Professor with Minjiang University, China. He has published more than 20 articles in international journals and conferences, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, International Journal of Computer Vision, Pattern Recognition, Pattern Recognition Letters, Computer Vision and Image Understanding, ICCV, ECCV, ACCV, AAAI, ICIP, and ICARCV. His research interests include machine learning, computer vision, and pattern recognition.



Yuanzheng Cai received the B.S. degree in software engineering from Fujian Normal University, Fuzhou, China, in 2010, the M.S. degree in computer science from Yunnan University, Kunming, China, in 2012, and the Ph.D. degree in artificial intelligence from Xiamen University, Xiamen, China, in 2016. He is currently a Lecturer with the College of Computer and Control Engineering, Minjiang University, Fuzhou. He is also with the College of Mathematics and Computer Science, Fuzhou University, Fuzhou, and NetDragon Inc., Fuzhou. His research interests include image/video retrieval, machine learning, and object recognition.