

Weakly But Deeply Supervised Occlusion-Reasoned Parametric Road Layouts

Buyu Liu¹ Bingbing Zhuang¹ Manmohan Chandraker^{1,2}
¹NEC Laboratories America ²UC San Diego

Abstract

We propose an end-to-end network that takes a single perspective RGB image of a complex road scene as input, to produce occlusion-reasoned layouts in perspective space as well as a parametric bird’s-eye-view (BEV) space. In contrast to prior works that require dense supervision such as semantic labels in perspective view, our method only requires human annotations for parametric attributes that are cheaper and less ambiguous to obtain. To solve this challenging task, our design is comprised of modules that incorporate inductive biases to learn occlusion-reasoning, geometric transformation and semantic abstraction, where each module may be supervised by appropriately transforming the parametric annotations. We demonstrate how our design choices and proposed deep supervision help achieve meaningful representations and accurate predictions. We validate our approach on two public datasets, KITTI and NuScenes, to achieve state-of-the-art results with considerably less human supervision.

1. Introduction

Understanding road layout from images is essential for real-world applications such as autonomous driving or path planning [5, 8, 13, 31], where, besides the usual perspective space outputs, top-view representations of geometry and semantics have been popular. Non-parametric representations such as pixel-level semantics [31] generally require labor-intensive and potentially ambiguous supervision in the top-view, for example, when dealing with occluded regions. On the other hand, parametric representations for top-view layouts are desirable for their interpretability, which is beneficial for higher-level reasoning and decision-making in downstream applications.

Parametric attributes such as presence of side roads or number of lanes may be easily annotated by humans given sensor inputs, and require less effort than pixel-level semantic annotations. However, besides parametric annotations¹ in

¹Parametric and attribute-level annotations are used interchangeably in our paper.

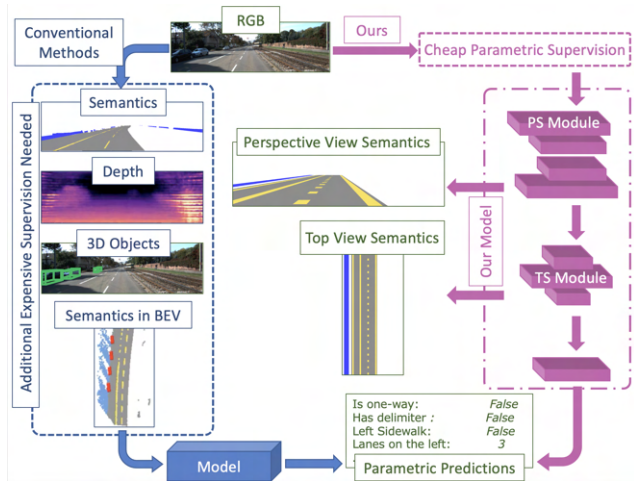


Figure 1. We propose an end-to-end model that inputs perspective image and outputs parametric layouts in top-view. Compared to existing methods, ours requires only the parametric layout annotations during training and achieves SOTA performance under complex road scenarios. Moreover, it generates occlusion-reasoned (see the predicted semantics on regions occluded by cars) pixel-level semantics in both perspective and top view.

the bird’s-eye-view (BEV), i.e. top-view, previous works that estimate parametric BEV layouts also require pixel-level supervision in perspective images [23, 46] or handle only very simple road layouts [35]. This paper seeks to obtain parametric BEV maps as well as pixel-level semantics in both the perspective and top views, but using only the cheaper parametric supervision on attributes.

While relying on cheap supervision only is undoubtedly a goal worth pursuing, removing the dense perspective supervision makes the problem harder. This is non-trivial, since there exists a large gap between sparse parametric supervision and dense pixel-level semantic supervision. To bridge the gap, one must reason about the underlying geometry to map the parametric supervision to top-view and get the correct semantics, even in occluded regions.

We address this challenge through two key insights. First, rather than directly regressing the parametric BEV layout from RGB image space, we introduce two intermediate steps – a perspective semantics (PS) module and a top-view se-

mantics (TS) module – to predict intermediate occlusion-reasoned per-pixel perspective and BEV layouts (Fig. 1). Second, to obtain supervision for PS/TS, a simple renderer can convert the parametric annotations to occlusion-reasoned per-pixel semantic annotations in both the BEV and perspective view, with the help of geometric transformation. This allows meaningful deep supervision [18, 19] of intermediate modules without additional annotation costs, thus weakly supervised. The weakly but deeply supervised PS and TS modules together lead to accurate parametric BEV layout by introducing inductive biases on the type of reasoning the network should perform, thereby facilitating complex tasks such as occlusion reasoning, geometric transformation and semantic abstraction that correspond to the parametric supervision.

The above insights make our method simple yet highly effective, even outperforming previous methods that rely on perspective-view dense supervision for semantic segmentation. We validate our choices through state-of-the-art (SOTA) accuracies on both KITTI [9] and NuScenes [28] datasets, achieving 47.3% and 13.0% F1 score. In extensive ablation experiments, we establish the value of the inductive biases introduced by the PS and TS modules, as well as the deep supervision through transformed parametric annotations.

To summarize, our key contributions are:

- An end-to-end model for occlusion-reasoned perspective and top-view parametric layout in complex scenes.
- Intermediate module design that incorporates inductive biases to learn occlusion-reasoning, geometric transformation and semantic abstraction.
- Deep supervision with cheap parametric annotations in top view only, rather than requiring additional expensive per-pixel labeling in either perspective or top view.
- State-of-the-art results on publicly available datasets.

2. Related Work

3D scene understanding on outdoor scenes is an important yet challenging task. Applications such as robot navigation [13], autonomous driving [8, 17], augmented reality [1] or real estate [24, 39] always require comprehensive understanding on given scenes.

Road Scene Understanding Scene understanding for outdoor scenarios is very challenging mainly due to the lack of strong priors. To this end, non-parametric approaches have been proposed [12, 40, 41], where layered representations [4, 48] are utilized to reason about the geometry as well as semantics in occluded areas. Other typical non-parametric representations in perspective view are joint pixel-level semantics and depth [21], pixel-level semantics and geometric

labels [11]. In contrast, parametric approaches provide abstract understanding, such as road scene attributes [8, 35] and graph-based representation [17]. Perhaps [23, 46] are the most recent works that are able to handle complex road layout, e.g. multiple lanes and different types of intersections. Our work follows the parametric representation proposed in these methods. Unlike [23, 46] that request additional information, e.g. models [15] pre-trained with dataset-specific per-pixel semantics, depth and 3D objects [9], to map semantics to top-view as pre-processing, our model is end-to-end trainable that directly takes RGB as input. More importantly, we exploit deep supervision [18, 19] by introducing meaningful intermediate modules (PS and TS), with which we are able to obtain occlusion-reasoned pixel-level semantics in both perspective and top-view without per-pixel human annotations. It is also beneficial in terms of improving final parametric layout predictions. Though focusing on single image for now, our model can be easily extend to video-version by introducing spatio-temporal graphical model [22, 46], LSTM [6, 38] or FTM [42, 49, 50]

Scene Understanding in Top-view Top-view representations [25, 27, 29, 37] can be more beneficial when occlusion relationships are desired, e.g., two objects cannot occupy the same position in top-view while they can potentially occlude each other in perspective view. Such intuition is widely exploited in 3D object localization literature [44] where camera to top-view projection is fulfilled with the help of depth estimation and 2D detection in perspective view. Although [32] proposes an end-to-end trainable model that explicitly exploits the perspective to top-view projection to perform 3D localization task, the performance of this method is not comparable to [44] due to the lack of explicit depth-aware re-projection. As for general scene understanding, the initial steps are taken in [34, 36]. However, due to the lack of ground truth, no quantitative evaluation is performed in [34]. More recent work [30, 31] extends [32] and predicts top-view semantic map from single monocular image or multiple streams of images. A graph like parametric representation is introduced in [2] for road layout estimation as well as oriented bounding boxes for road participants. However, such representation misses important semantics such as crosswalk, sidewalk and lane directions. And it further requires HD-map, GPS and human annotations to train such a model. In contrast to non-parametric approaches [26, 29–31, 47] that require expensive per-pixel supervision in top or perspective view and focus on predicting semantics on visible regions, our method aims to predict parametric layouts in BEV and also provides occlusion-aware non-parametric representations in both BEV and perspective view as by-products. All these meaningful representations are obtained without per-pixel human annotations but relying on cheap parametric annotations.

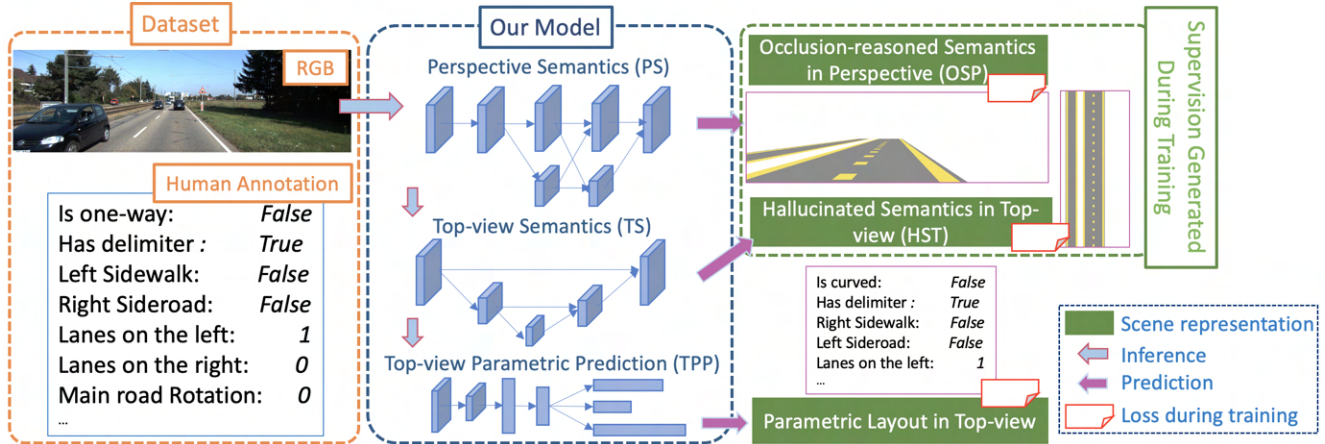


Figure 2. **Overview of our proposed framework:** Taking a single RGB as input, our model predicts (1) occlusion-reasoned semantics in perspective view, (2) hallucinated semantics in top-view and (3) parametric layout predictions in top-view, with only *attribute-level* annotations in top-view. This is achieved with multiple intermediate modules and deeply supervised training.

3. Our Framework

Our model consists of three modules. (1) The perspective semantics (PS) module inputs the RGB image and outputs the occlusion-reasoned pixel-level semantics in perspective view (OSP). (2) The top-view semantics (TS) module projects OSP into top-view and learns to hallucinate or complete pixel-level top-view semantics on out-of-view as well as noisy regions, which we refer to as hallucinated semantics in top-view (HST). (3) The top-view parametric prediction (TPP) module takes the HST and predicts road layout related attributes in top-view. Fig. 2 gives an overview of the proposed method. Network architectures are borrowed from [33, 43, 46] and described in Sec. 4 and supplementary. We focus in this section on describing our main contributions that allow effectively exploiting weak supervision with cheap parametric-level human annotations. We detail each module in Sec. 3.1, the training process and the generation of intermediate pixel-level semantic annotations in Sec. 3.2.

3.1. Full Model

Consider a dataset $\mathcal{D} = \{I, \Theta\}_{i=1}^N$ of N samples, where $I \in \mathbb{R}^{H \times W \times 3}$ are RGB perspective images and Θ denote the corresponding scene attributes obtained from human annotations. We further generate x^p, x automatically for each sample where $x^p \in \mathbb{R}^{H \times W \times (C+1)}$ denotes semantic segmentation map in perspective view and $x \in \mathbb{R}^{h \times w \times (C+1)}$ denotes top-view semantics. $C = 4$ denotes the number of layout categories (“road”, “sidewalk”, “lane boundaries”, “crosswalks”) and we also include one foreground class. We refer the readers to Sec. 3.2 for more details about the data generation process. Our full model is defined as:

$$\Theta = f^{\text{full}}(I) = (f^{\text{tpp}} \circ f^{\text{ts}} \circ f^{\text{ps}})(I), \quad (1)$$

where \circ defines a function composition. f^{ps} , f^{ts} and f^{tpp} correspond to our three modules PS, TS, and TPP.

Perspective semantics module The PS module predicts per-pixel occlusion-reasoned semantics in the perspective view (OSP). Unlike traditional semantic segmentation models (e.g. [3, 43]) that predict semantics on visible pixels only, our module focuses on predicting both visible and occluded layout classes (See Fig. 3(d)). Such occlusion reasoning is also demonstrated in Fig. 3(b) and (c). As shown, we aim to predict road semantics in the top-view despite that they are occluded, e.g. by cars or buildings, in the perspective view.

Compared to conventional semantic segmentation problem, ours is more challenging in terms of both data and model training. As for data, the semantic ground-truth on occluded regions can be ambiguous, hence difficult and time-consuming to annotate accurately in pixel-level. For instance, it takes more than 20 minutes to annotate only the visible regions on KITTI images while in comparison, parametric annotation in BEV takes about 20s for an image [45]. We refer the readers to Sec. 4 and supplementary for annotation details. For model training, the PS module predicts semantics in invisible/occluded regions, which again requires dealing with ambiguity. For instance, regions occluded by a foreground instance, e.g. building, can be either another building or road. This requires the module to learn to predict semantics with contextual cues rather than fully relying on local visible information.

Formally, given an image I , the PS module outputs x^p encoding the probability of each pixel belonging to a specific category:

$$x^p = f^{\text{ps}}(I). \quad (2)$$

Top-view semantics module Our second module, i.e. the top-view semantics module, takes as input the OSP and

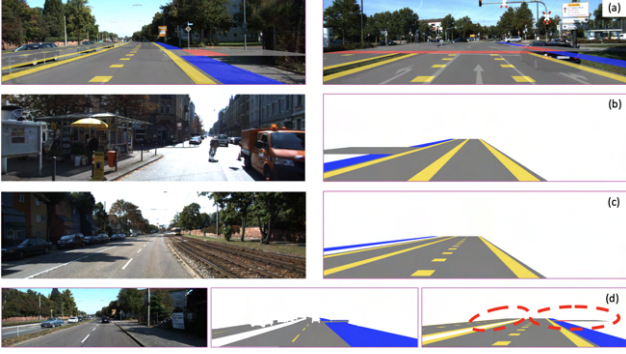


Figure 3. We overlay output and input of PS module in (a). Examples of inputs (left) and target outputs (right) of PS module are provided in (b) and (c). PS module aims to predict both visible and occluded background classes. (d) demonstrates the input image, semantics of visible regions and our target output from left to right. And we highlight the occluded regions in red.

learns to explicitly project the semantics in perspective view to top-view. Given camera intrinsics, the projection could be done if the depth estimation is available, say, via a depth network. However, standard single image depth networks (e.g. [7, 10, 21]) typically do not reason about depth in occluded regions, which is nevertheless required for our occlusion-aware projection. In addition, resolution is low for distant regions and thus may lead to sparser/noisier semantics in top-view. Lastly, top-view semantics on close-by regions can be incomplete due to limited field of view. Instead, we propose a two-step projection through an initial geometric transformation f^{trans} and a learned hallucination module f^{halln} :

$$x = f^{\text{ts}}(x^p) = (f^{\text{halln}} \circ f^{\text{trans}})(x^p). \quad (3)$$

Transformation module. In view of these issues, we first make use of the prior that the road forms nearly a plane, which facilitates an initial projection without requiring depth estimation. We assume known camera intrinsics and extrinsics w.r.t. the ground plane; this is a mild assumption since they could be obtained via calibration [14] in advance. As such, it is well-known that one can back-project each pixel in the perspective view to the BEV view and vice versa [14].

Hallucination module. After the transformation module that maps the OSP to top-view, the hallucination module then learns to predict the unseen far away regions as well as recover the noisy semantics with contextual information in top-view. Note that our input and output of hallucination module are both of the size $h \times w \times (C+1)$. Fig. 4 visualizes two sets of inputs and outputs of this module. Compared to inputs generated with ground-truth OSP, the target HST improves at far away (right) regions as well as close-by areas where predictions are sparse (left).

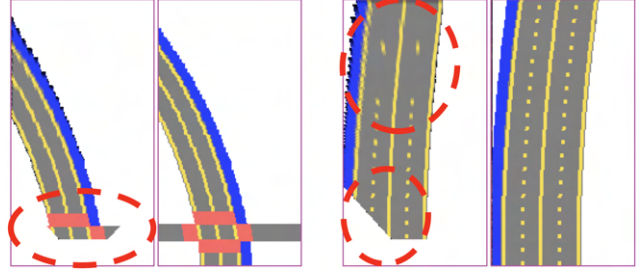


Figure 4. Two sets of examples for input and output of hallucination module. Our module aims to recover the far away sparse semantics (right) and hallucinate close-by areas with limited view (left).

Top-view parametric prediction module Given the hallucinated semantics in top-view (HST), our next step is to predict the layout attributes through the top-view parametric prediction (TPP) module that maps the HST x into the scene model parameters Θ . As aforementioned, we follow the attribute definitions in [23, 46]. Our Θ consists of three groups: Θ_b for 14 binary, Θ_m for 2 multi-class and Θ_c for 10 continuous attributes of the scene model, respectively. Binary attributes consist of information such as whether the road is one-way or not. Number of lanes on the left hand-side of the ego car is an example of multi-class attributes and distance to right side-road can be one of the continuous attributes. More details can be found in supplementary materials. Our TPP module is defined as:

$$\Theta = f^{\text{tpp}}(x) = (f \circ g)(x), \quad (4)$$

where f and g are respectively multi-layer perceptron (MLP) and convolutional neural networks. Note that similar to [46], this module is also able to exploit rich simulated data during training, but we leave this extension to future work.

3.2. Model Training

Following our above description of intermediate modules assuming supervision available, we describe in this section the generation of such supervision with only annotations for parametric layout Θ , as well as deep supervision training. Instead of training the full model in an end-to-end manner from scratch, we adopt a multi-stage training protocol. We first pretrain all three modules, and then jointly train the full model in an end-to-end manner. Empirically, end-to-end training provides a 1% performance improvement in our experiments. Our full loss function \mathcal{L} is defined as:

$$\mathcal{L} = \lambda \mathcal{L}^{\text{tpp}} + \gamma \mathcal{L}^{\text{ts}} + \beta \mathcal{L}^{\text{ps}}, \quad (5)$$

where λ , γ and β are the weights for each module.



Figure 5. Examples of rendered ground-truth for TS module. From left to right: RGB, parametric human annotations and rendered pixel-level semantics in top-view.

Top-view parametric prediction module Since Θ and I are already available, we define the loss function of TPP as:

$$\mathcal{L}^{\text{TPP}} = \sum_{i=1}^N \text{BCE}(\Theta_{b,i}, \eta_{b,i}) + \text{CE}(\Theta_{m,i}, \eta_{m,i}) + \ell_1(\Theta_{c,i}, \eta_{c,i}), \quad (6)$$

where (B)CE is the (binary) cross-entropy loss and ℓ_1 denotes L1 loss. $\{\Theta, \eta\}_{\cdot, i}$ denotes the i -th sample in the data set. For regression, we discretize continuous variables into 100 bins by convolving a dirac delta function centered at Θ_c with a Gaussian of fixed variance.

Top-view semantics module Unlike the straightforward design in parametric space, our TS module requires per-pixel supervision in top-view. To this end, we propose to exploit a rendering function that generates pixel-wise semantics from parametric annotations. Specifically, for each Θ , we render a map x . Some examples of our paired $\{x, \Theta\}$ are in Fig. 5, which shows that our rendered x^p accurately reflects the layout of the road in top-view. Since we only need on parametric abstractions, our renderer can be implemented using simple Python code, rather than the complex machinery of physics-based image renderers. We refer the readers to supplementary materials for more details on our renderer and the generation process. The loss function for TS module is defined as:

$$\mathcal{L}^{\text{TS}} = \sum_{i=1}^N \text{CE}(x_i, \hat{x}_i) \quad (7)$$

where \hat{x}_i and x_i denotes the predictions and the rendered ground-truth of the top-view semantics of i -th sample in \mathcal{D} .

Perspective semantics module Obtaining the top-view semantics x , we can project [14] it to perspective view with camera parameters as well as plane assumption. We demonstrate the effectiveness of our projection in Fig. 3(a). Similarly, the loss function for the PS module is defined as:

$$\mathcal{L}^{\text{PS}} = \sum_{i=1}^N \text{CE}(x_i^p, \hat{x}_i^p) \quad (8)$$

where \hat{x}_i^p and x_i^p denotes our predictions and the back-projected ground-truth of the perspective semantics of sample i in \mathcal{D} .

4. Experiments

Datasets and model details We validate our ideas on KITTI [9] and NuScenes [28], utilizing the annotation and data split in [23]. Please refer to [23] and our supplementary for details in road layout attributes annotation. h and w are set to 256 and 128, presenting a $60\text{m} \times 30\text{m}$ space in real world. Camera parameters are available in the original datasets through calibration. Weights (λ, γ, β) are set experimentally on validation set. As for f^{ps} , we use HRNetV2-W18 [43] as the backbone as it achieves very good trade-offs between accuracy and efficiency. As for f^{halln} , we utilize a shallower version of [33], e.g. 5-layer encoder and decoder. Finally, f is implemented as a multi-task network with three separate predictions η_b, η_m and η_c for each of the parameter groups Θ_b, Θ_m and Θ_c of the scene model. And g is introduced for feature extraction. Note that our method does not depend on the specific details of these sub-modules but is generally applicable if this three-stage architecture holds.

Cost for parametric annotations We summarize the annotation time for each type of supervision in Tab. 2. Unlike non-parametric annotations such as pixel-level semantics that require several dozens of minutes per frame, our parametric annotations require less than a minute per frame. Moreover, this time is heavily amortized across a video sequence to just around 20 seconds on the KITTI dataset, since parametric attributes change predictably across consecutive frames. Binary and multiclass attributes (such as presence of side-road, or number of lanes) change less frequently and their annotations can often be inherited from previous frames. Further, continuous attributes (such as distance to intersection) typically change smoothly across frames, which facilitates annotation. We refer the readers to supplementary materials for more details.

Evaluation metrics Since our output space Θ consists of three types of predictions and involves both discrete and continuous variables, we follow the metrics in [23, 46]. Specifically, as for binary variables Θ_b and multi-class variables Θ_m , the prediction accuracy is defined as $\text{Accu.-Bi} = \frac{1}{14} \sum_{k=1}^{14} [p_k = \Theta_{bk}]$ and $\text{Accu.-Mc} = \frac{1}{2} \sum_{k=1}^2 [p_k = \Theta_{mk}]$. We further report the F1 score on Θ_b to have a better idea about the overall performance given the observation that the binary classes are extremely biased. Formally, $\text{F1} = \frac{1}{14} \sum_{k=1}^{14} 2 \times \frac{p_k \times r_k}{p_k + r_k}$, where p_k and r_k are the precision and recall rate on Θ_{bk} . For continuous variables, we report the mean square error (MSE).

Method	Supervision Required					KITTI [9]			
	Parametric	Depth	Semantics	Simulated	Video+Object	Accu.-Bi. \uparrow	Accu.-Mc. \uparrow	MSE \downarrow	F1 \uparrow
RGB [15, 35]	✓					.811	.778	.230	.176
RGB [15, 35]+D	✓	✓				.818	.819	.154	.109
BEV [34]	✓	✓	✓			.820	.797	.141	.324
H-BEV+DA [46]	✓	✓	✓	✓		.834	.831	.134	.435
BEV-J-O [23]	✓	✓	✓		✓	.831	.837	.142	.494
Ours	✓					<u>.833</u>	<u>.832</u>	<u>.140</u>	<u>.473</u>

Table 1. Performances on single image based road layout prediction on KITTI. we observe that our method outperforms *RGB* when having the same model setting. In addition, our results are comparable to other SOTA (*H-BEV+DA* and *BEV-J-O*) but with far less human annotations required.

Time	Binary	Multiclass	Continuous	Total
Random images	24.3	5.1	25.7	55.1
Video frames	20.2			

Table 2. Average annotation time (sec.) on KITTI dataset.

Apart from parametric predictions, our model also outputs intermediate representations, e.g. OSP and HST. We further report the IoU as well as the accuracy for these two semantic segmentation tasks. Please note that human annotated OSP and HST are not available on either dataset in practice. Thus, we report our performance by comparing our predictions with rendered semantics x and x^p instead.

4.1. Evaluations of Parametric Road Layout

Baselines We choose several appropriate baselines as presented in [23, 46].:

- **RGB (*RGB*)**: A ResNet-101 [16, 35] backbone is introduced and trained on the manually-annotated ground truth. Note that this setup is the **only** one that directly comparable to ours as it requires only the parametric annotations as ground-truth.
- **RGB+Depth (*RGB + D*)**: Same as *RGB* but with the additional task of monocular pixel-wise depth prediction [16]. In contrast, we do not require dense depth information.
- **BEV (*BEV*)**: *BEV* uses the output of [34], which is a top-view semantic map. To obtain such map, additional pixel-level semantic annotation and depth supervision are required in perspective space. Though more recent approaches [26, 30] are also able to output semantics in top-view, they miss importance semantics such as lane boundary or crosswalk thus are not desired as *BEV* baselines.

We also report the performance of SOTA methods for single image top-view layout prediction, or **H-BEV-DA** [46] and **BEV-J-O** [23]. Please note that both of them require far more human annotations compared to our method. We refer the readers to supplementary for more details for all baselines.

Quantitative results Tab. 1 summarizes our main results on KITTI [9]. First of all, if we compare only to method with the same setting, or *RGB*, our method outperforms it with a large margin, which indicates the effectiveness of introducing PS and TS as intermediate modules. Furthermore, compared to *RGB + D* method that introduces depth channel, or even the *BEV* that further requires thousands of human labelled semantic segmentation images in perspective space, our method achieves better results, which is of significance given that our method requires far less human annotations. Note that both *H-BEV-DA* and *BEV-J-O* are based on *BEV* but require even more human annotations. By comparing to *H-BEV-DA* that further exploits additional simulated data and *BEV-J-O* that requires the 3D object information as well as entire video sequence as input, we can see that our method achieves comparable results with far less human annotations.

We further report results on NuScenes [28] in Tab. 3. Our method outperforms *RGB* significantly. It also outperforms [23, 46] with far less human annotations required.

Qualitative results We demonstrate some qualitative results in Fig. 6. Note that in KITTI test sequences, rather than the road being occluded by cars driving in front, significant occlusions happen between parked cars and road/sidewalk, or between foreground classes, e.g. buildings or trees, and curved road or sideroad. As observed in this figure, our model is able to output satisfactory results on all three representations. We are able to handle complex road layout such as arbitrary number of lanes with heavy occlusions. Again, please note that OSP and HST are obtained without per-pixel human annotations. Our final layout prediction is also better than [23]. We further visualize our final results on NuScenes in Fig. 7. It shows that our model is able to handle various road layouts. We refer readers to supplementary material for more qualitative results.

4.2. Ablation Study

To demonstrate the effectiveness of intermediate modules as well as deep supervision, we further conduct exper-

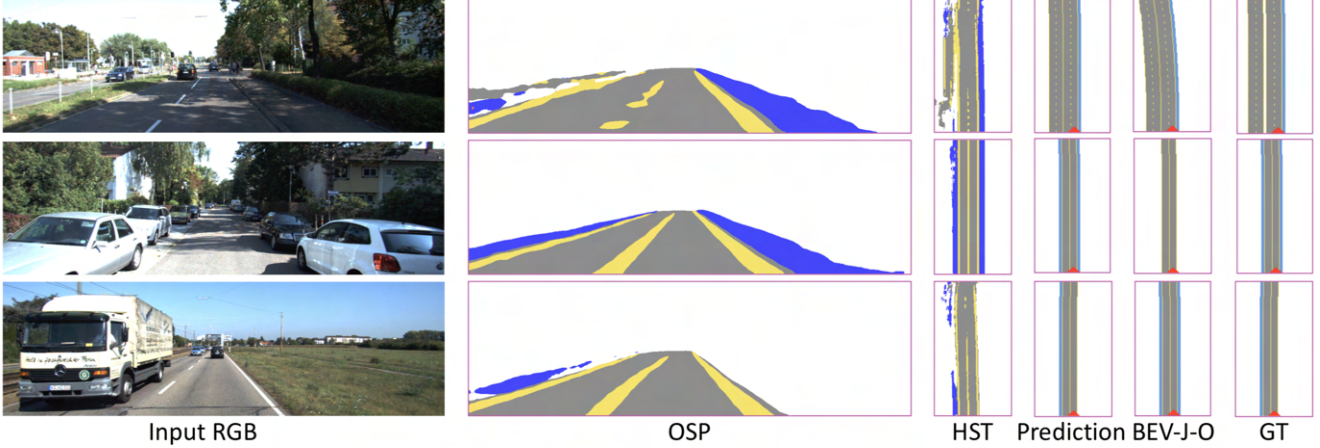


Figure 6. Full predictions of our proposed model. From left to right: input RGB, OSP, HST, image rendered from parametric predictions, results from [23] and image rendered from ground-truth attributes.

Method	Accu.-Bi. \uparrow	Accu.-Mc. \uparrow	MSE \downarrow	F1 \uparrow
RGB [15, 35]	.850	.503	.084	.109
BEV [46]	.846	.485	.073	.101
H-BEV+DA [46]+GM	.877	.496	.032	.125
BEV-J-O [23]	.858	.543	.027	.128
Ours	<u>.875</u>	.560	.023	.130

Table 3. Results on NuScenes dataset. We observe that our method beats *RGB* significantly when having the same model setting. Meanwhile, it also outperforms *H-BEV+DA* and *BEV-J-O* with far less human annotations required.

Method	Module				KITTI [9]			
	f^{ps}	f^{trans}	f^{halln}	f^{tpp}	Accu.-Bi. \uparrow	Accu.-Mc. \uparrow	MSE \downarrow	F1 \uparrow
RGB				\checkmark	.811	.778	.230	.176
RGB+PS	\checkmark			\checkmark	.822	.827	.159	.425
RGB+PS+T	\checkmark	\checkmark		\checkmark	.826	.829	.144	.441
Ours	\checkmark	\checkmark	\checkmark	\checkmark	.833	.832	.140	.473

Table 4. Ablation study on single image based road layout prediction on KITTI. Note that all these methods share the same amount of human annotations. We can see that our introduced PS and TS modules, on the one hand, provide meaningful intermediate representations at no additional costs. On the other hand, they also prove to be beneficial individually for the final parametric prediction task.

iments on incrementally adding modules without any module. *RGB+PS* directly predicts parametric predictions. Formally, *RGB+PS* is fo



$$\Theta = f^{rbgp}(I) = (f^{tpp} \circ f^{ps})(I), \quad (9)$$

Similarly, *RGB+PS+T* is formulated as:

$$\Theta = f^{rbgpf}(I) = (f^{tpp} \circ f^{trans} \circ f^{ps})(I), \quad (10)$$

We report the quantitative results in Tab. 4. The results show that first of all, comparing the *RGB* to *RGB+PS*, perspective representation, or the OSP, is beneficial for improving final parametric predictions. Secondly, the performance gap between *RGB+PS+T* and *RGB+PS* demonstrates the effectiveness of introducing top-view semantics as intermediate representation. Finally, by comparing the full model with *RGB+PS+T*, we can tell that the hallucination module is also critical for layout prediction task.

Occlusion study Here, we study performance against increasing number of objects in the scene, indicating increasingly severe occlusions. Since [9, 45] do not provide pixel-

Figure 7. Examples on NuScenes dataset. Left:input RGB Right: rendered BEV semantics from our prediction.

level semantic ground-truths on our test sequences, to analyze our ability to handle occlusions, we instead report the average image-level IoU on four classes against number of foreground objects with respect to rendered ground truth x in Tab. 6, with objects detected by Stereo-RCNN [20]. As one can see, our method outperforms the state-of-the-art consistently, with increasing gap when having more objects. Please note that our model is single-image based so we handle all objects, no matter they are moving or not, in the same manner.

4.3. Evaluations of Intermediate Representations

Apart from requiring less annotation while maintaining comparable performance, another advantage of the proposed method is being able to provide meaningful pixel-level in-

Data	Representation		KITTI [9]											
	OSP	HST	Road		Land Boundary		Sidewalk		Crosswalk		Foreground		Average	
			Accu.	IoU	Accu.	IoU	Accu.	IoU	Accu.	IoU	Accu.	IoU	Accu.	IoU
RGB+PS	✓		.689	.563	.365	.214	.226	.126	.010	.007	.954	.878	.449	.358
Ours	✓		.700	.605	.403	.272	.255	.147	.042	.033	.962	.883	.472	.388
		✓	.605	.461	.272	.197	.167	.102	.038	.032	.868	.651	.390	.289

Table 5. Intermediate results on KITTI. We report both IoU and accuracy for each semantic category. Compared to *RGB+PS*, our method achieves better performance in terms of *OSP* with the help of end-to-end training. Our method further provides meaningful *HST* results.

Obj.	0	1	2	3	4	5	6	7	8	Avg.
[46]	.67	.78	.67	.64	.61	.45	.48	.37	.35	.45
Ours	.78	.81	.72	.69	.67	.48	.50	.40	.35	.50
Considering Road Class Only										
[46]	.85	.74	.85	.79	.67	.63	.59	.57	.37	.63
Ours	.86	.91	.86	.83	.70	.70	.61	.61	.50	.70

Table 6. Average per-image IoU w.r.t. number of road participants.

intermediate representations, OSP and HST, as by-products. To demonstrate that these intermediate representations are indeed semantically useful for downstream tasks, we study their IoU as well as accuracy score, as an indication for their performance. Please note that compared to existing work that requires dense and time-consuming pixel-wise human annotation, ours only requires cheap parametric human annotations and produces pixel-level occlusion-reasoned semantic segmentation in perspective and top-view.

As shown in Tab. 5, our method is able to provide multiple meaningful intermediate representations. Also, our deep supervision proves to be beneficial in an end-to-end manner, which can be observed from the performance gap on OSP between *RGB+PS* and our full model. In addition, our model also achieves reasonably good performance on HST. As a reference, [31], which aims to predict pixel-level semantics of visible regions in top-view with perspective images as input, reports about 63.0% IoU for drivable category on two different datasets.

However, please note that [31] requires pixel-level dense annotations in top-view during training and the predictions are not occlusion-reasoned. We further visualize quantitative results in Fig. 8 and Fig. 9. As can be seen, our method obtain high quality semantics in both perspective and top-view despite occlusions.

5. Conclusion

We propose a novel end-to-end model that inputs single RGB perspective image and outputs multi-aspect representations for road layout, including top-view parametric predictions, OSP and HST. Specifically, we introduce two intermediate modules and exploit deep supervision to learn inductive biases in occlusion-reasoning, geometric transformation and semantic abstraction. We demonstrate the

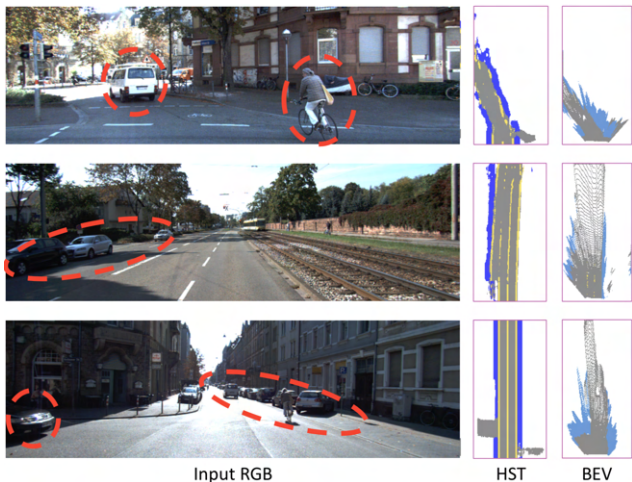


Figure 8. We demonstrate input RGB, predicted HST as well as BEV of [46], which is trained with thousands of pixel-level annotated images and LiDAR images. As can be seen in these examples, our model is able to hallucinate far away regions in a realistic manner, even on curved road, with *NO* pixel-level human annotations.



Figure 9. Input image, generated ground-truth pixel-level semantics and predicted semantics from top to bottom row. Our model is able to predict the semantics quite well despite occlusions.

effectiveness of our proposed method as well as intermediate modules on publicly available datasets and demonstrate that we can achieve SOTA performance with less human annotations.

References

- [1] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3D Semantic Parsing of Large-Scale Indoor Spaces. In *CVPR*, 2016. 2
- [2] Yigit Baran Can, Alexander Liniger, Danda Pani Paudel, and Luc Van Gool. Structured bird’s-eye-view traffic scene understanding from on board images. In *ICCV*, 2021. 2
- [3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 3
- [4] Helisa Dhama, Nassir Navab, and Federico Tombari. Object-driven multi-layer scene decomposition from a single image. In *ICCV*, 2019. 2
- [5] Vikas Dhiman, Quoc-Huy Tran, Jason J. Corso, and Manmohan Chandraker. A Continuous Occlusion Model for Road Scene Understanding. In *CVPR*, 2016. 1
- [6] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *CVPR*, 2017. 2
- [7] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016. 4
- [8] Andreas Geiger, Martin Lauer, Christian Wojek, Christoph Stiller, and Raquel Urtasun. 3D Traffic Scene Understanding from Movable Platforms. *PAMI*, 2014. 1, 2
- [9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets Robotics: The KITTI Dataset. *IJRR*, 2013. 2, 5, 6, 7, 8
- [10] Clément Godard, Oisín Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 4
- [11] Stephen Gould, Richard Fulton, and Daphne Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009. 2
- [12] Ruiqi Guo and Derek Hoiem. Beyond the line of sight: labeling the underlying surfaces. In *ECCV*, 2012. 2
- [13] Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive Mapping and Planning for Visual Navigation. In *CVPR*, 2017. 1, 2
- [14] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4, 5
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 6, 7
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 6
- [17] Lars Kunze, Tom Bruls, Tarlan Suleymanov, and Paul Newman. Reading between the Lanes: Road Layout Reconstruction from Partially Segmented Scenes. In *International Conference on Intelligent Transportation Systems (ITSC)*, 2018. 2
- [18] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-Supervised Nets. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 2015. 2
- [19] Chi Li, M. Zeeshan Zia, Quoc-Huy Tran, Xiang Yu, Gregory D. Hager, and Manmohan Chandraker. Deep supervision with shape concepts for occlusion-aware 3d object parsing. In *CVPR*, 2017. 2
- [20] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *CVPR*, 2019. 7
- [21] Beyang Liu, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels. In *CVPR*, 2010. 2, 4
- [22] Buyu Liu and Xuming He. Multiclass semantic video segmentation with object-level active inference. In *CVPR*, 2015. 2
- [23] Buyu Liu, Bingbing Zhuang, Samuel Schuster, Pan Ji, and Manmohan Chandraker. Understanding road layout from videos as a whole. In *CVPR*, 2020. 1, 2, 4, 5, 6, 7
- [24] Chenxi Liu, Alexander G. Schwing, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Rent3D: Floor-Plan Priors for Monocular Layout Estimation. In *CVPR*, 2015. 2
- [25] Chenyang Lu, Marinus Jacobus Gerardus van de Molengraft, and Gijs Dubbelman. Monocular Semantic Occupancy Grid Mapping With Convolutional Variational Encoder-Decoder Networks. *IEEE Robotics and Automation Letters*, 2019. 2
- [26] Kaustubh Mani, Swapnil Daga, Shubhika Garg, Sai Shankar Narasimhan, Madhava Krishna, and Krishna Murthy Jatavallabhula. Monolayout: Amodal scene layout from a single image. In *WACV*, 2020. 2, 6
- [27] Kaustubh Mani, N Sai Shankar, Krishna Murthy Jatavallabhula, and K Madhava Krishna. Autolay: Benchmarking amodal layout estimation for autonomous driving. In *IROS*, 2020. 2
- [28] NuTonomy. The NuScenes data set. <https://www.nuscenes.org>, 2018. 2, 5, 6
- [29] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou. Cross-view semantic segmentation for sensing surroundings. *IEEE Robotics and Automation Letters*, 2020. 2
- [30] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. *arXiv preprint arXiv:2008.05711*, 2020. 2, 6
- [31] Thomas Roddick and Roberto Cipolla. Predicting semantic map representations from images using pyramid occupancy networks. In *CVPR*, 2020. 1, 2, 8
- [32] Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection. *arXiv preprint arXiv:1811.08188*, 2018. 2
- [33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015. 3, 5
- [34] Samuel Schuster, Menghua Zhai, Nathan Jacobs, and Manmohan Chandraker. Learning to Look around Objects for Top-View Representations of Outdoor Scenes. In *ECCV*, 2018. 2, 6
- [35] Ari Seff and Jianxiong Xiao. Learning from Maps: Visual Common Sense for Autonomous Driving. *arXiv:1611.08583*, 2016. 1, 2, 6, 7

- [36] Sunando Sengupta, Paul Sturgess, Ľubor Ladický, and Philip H. S. Torr. Automatic Dense Visual Semantic Mapping from Street-Level Imagery. In *IROS*, 2012. [2](#)
- [37] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geolocalization. *NeurIPS*, 2019. [2](#)
- [38] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. [2](#)
- [39] Shuran Song, Andy Zeng, Angel X. Chang, Manolis Savva, Silvio Savarese, and Thomas Funkhouser. Im2Pano3D: Extrapolating 360 Structure and Semantics Beyond the Field of View. In *CVPR*, 2018. [2](#)
- [40] Joseph Tighe, Marc Niethammer, and Svetlana Lazebnik. Scene Parsing with Object Instances and Occlusion Ordering. In *CVPR*, June 2014. [2](#)
- [41] Shubham Tulsiani, Richard Tucker, and Noah Snavely. Layer-structured 3D Scene Inference via View Synthesis. In *ECCV*, 2018. [2](#)
- [42] Tuan-Hung Vu, Wongun Choi, Samuel Schulter, and Manmohan Chandraker. Memory warps for learning long-term online video representations. *arXiv preprint arXiv:1803.10861*, 2018. [2](#)
- [43] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *PAMI*, 2020. [3](#), [5](#)
- [44] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*, 2019. [2](#)
- [45] Ziyang Wang, Buyu Liu, Samuel Schulter, and Manmohan Chandraker. A dataset for high-level 3d scene understanding of complex road scenes in the top-view. In *CVPR Workshop*, 2019. [3](#), [7](#)
- [46] Ziyang Wang, Buyu Liu, Samuel Schulter, and Manmohan Chandraker. A parametric top-view representation of complex road scenes. In *CVPR*, 2019. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [47] Weixiang Yang, Qi Li, Wenxi Liu, Yuanlong Yu, Yuexin Ma, Shengfeng He, and Jia Pan. Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation. In *CVPR*, 2021. [2](#)
- [48] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion. In *CVPR*, 2020. [2](#)
- [49] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *ICCV*, 2017. [2](#)
- [50] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *CVPR*, 2017. [2](#)