A Parametric Top-View Representation of Complex Road Scenes

Ziyan Wang[1*]   Buyu Liu[2]   Samuel Schulter[2]   Manmohan Chandraker[2,3]

[1]Carnegie Mellon University   [2]NEC Laboratories America   [3]UC San Diego

Abstract

In this paper, we address the problem of inferring the layout of complex road scenes given a single camera as input. To achieve that, we first propose a novel parameterized model of road layouts in a top-view representation, which is not only intuitive for human visualization but also provides an interpretable interface for higher-level decision making. Moreover, the design of our top-view scene model allows for efficient sampling and thus generation of large-scale simulated data, which we leverage to train a deep neural network to infer our scene model's parameters. Specifically, our proposed training procedure uses supervised domain-adaptation techniques to incorporate both simulated as well as manually annotated data. Finally, we design a Conditional Random Field (CRF) that enforces coherent predictions for a single frame and encourages temporal smoothness among video frames. Experiments on two public data sets show that: (1) Our parametric top-view model is representative enough to describe complex road scenes; (2) The proposed method outperforms baselines trained on manually-annotated or simulated data only, thus getting the best of both; (3) Our CRF is able to generate temporally smoothed while semantically meaningful results.

1. Introduction

Understanding complex layouts of the 3D world is a crucial ability for applications like robot navigation, driver assistance systems or autonomous driving. Recent success in deep learning-based perception systems enables pixel-accurate semantic segmentation [3, 4, 33] and (monocular) depth estimation [9, 15, 32] in the perspective view of the scene. Other works like [10, 23, 25] go further and reason about occlusions and build better representations for 3D scene understanding. The representation in these works, however, is typically non-parametric, *i.e.*, it provides a semantic label for a 2D/3D point of the scene, which makes higher-level reasoning hard for downstream applications.

In this work, we focus on understanding driving scenarios and propose a rich *parameterized* model describing com-
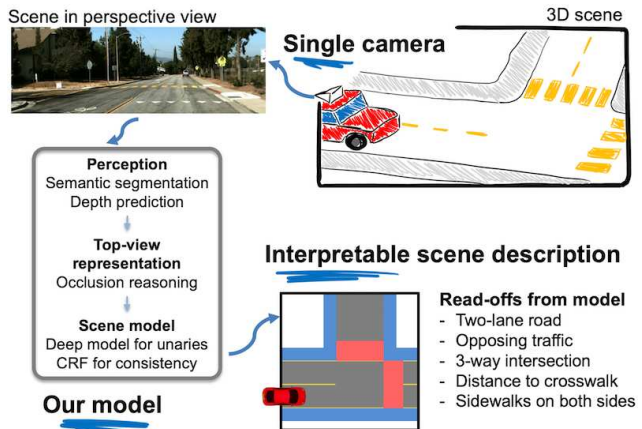

Figure 1: Our goal is to infer the layout of complex driving scenes from a single camera. Given a perspective image (top left) that captures a 3D scene, we predict a rich and interpretable scene description (bottom right), which represents the scene in an occlusion-reasoned semantic top-view.

plex road layouts in a top-view representation (Fig. 1 and Sec. 3.1). The parameters of our model describe important scene attributes like the number and width of lanes, and the existence of, and distance to various types of intersections, crosswalks and sidewalks. An explicit model of such parameters is beneficial for higher-level modeling and decision making as it provides a tangible interface to the real world. In contrast to prior art [7, 14, 17, 23, 24, 25], our proposed scene model is richer, fully parameterized and can be inferred from a single camera input with a combination of deep neural networks and a graphical model.

However, training deep neural networks requires large amounts of training data. Although annotating the scene attributes of our model for real RGB images is possible, it is also costly to do at a large-scale and, more importantly, extremely difficult for certain scene attributes. While the existence of a crosswalk is a binary attribute and is easy to annotate, annotating the exact width of a side road requires the knowledge of scene geometry, which is hard when only given a perspective RGB image. We thus propose to leverage simulated data. However, in contrast to rendering photo-realistic RGB images, which is a difficult and time-

consuming task [20, 21], we propose a scene model that allows for efficient sampling and render semantic top-view representations that obviate expensive illumination modeling or occlusion reasoning.

Given simulated data with accurate and complete annotations, as well as real images with potentially noisy and incomplete annotations, we propose a hybrid training procedure leveraging both sources of information. Specifically, our neural network design involves (i) a feature extractor that aims to leverage information from both domains, simulated and real semantic top-views from [23] (see Fig. 4), and (ii) a domain-agnostic classifier of scene parameters. At test time, we convert a perspective RGB image into a semantic top-view representation using [23] and predict our scene model's parameters. Given the individual scene parameter predictions, we further design a graphical model (Sec. 3.4) that captures dependencies among scene attributes in single images and enforces temporal consistency across a sequence of frames. We validate our idea on two public driving data sets, KITTI [8] and NuScenes [18] (Sec. 4). The results demonstrate the effectiveness of the top-view representation, the hybrid training procedure with real and simulated data, and the importance of the graphical model for coherent and consistent outputs. To summarize, our key contributions are:

- A **novel parametric and interpretable model** of complex driving scenes in a top-view representation.
- A neural network that (i) predicts the parameters from a single camera and (ii) is designed to enable a **hybrid training approach from both real and synthetic data**.
- A graphical model that ensures **coherent and temporally consistent** scene description outputs.
- **New annotations** of our scene attributes for the KITTI [8] and NuScenes [18] data sets[1].

2. Related Work

3D scene understanding is an important task in computer vision with many applications for robot navigation [11], self-driving [7, 14], augmented reality [1] or real estate [16, 27].

Scene understanding: Explicit modeling of the scene is frequently done for indoor applications where strong priors about the layout of rooms can be leveraged [1, 16, 26]. Non-parametric approaches are more common for outdoor scenarios because the layout is typically more complex and harder to capture in a coherent model, with occlusion reasoning often being a primary focus. Due to the natural ability to reflect orders, layered representations [10, 29, 31] have been utilized in scene understanding to reason about geometry and semantics in occluded areas. However, such intermediate representations are not desired for applications where distance information is required. A top-view representation [23, 25], in contrast, is a more detailed description

---

for 3D scene understanding. Our work follows such a top-view representation and aims to infer a parametric model of complex outdoor driving scenes from a single input image.

A few parametric models have been proposed for outdoor environments too. Seff and Xiao [24] present a neural network that directly predicts scene attributes from a single RGB image. Although those attributes are automatically acquired from OpenStreetMaps [19], they are not rich enough to fully describe complex road scenes, e.g. curved road with side-roads. A richer model that is capable of handling complex intersections with traffic participants is proposed by Geiger *et al*. [7]. To this end, they propose to utilize multiple modalities such as vehicle tracklets, vanishing points and scene flow. Different from their work, we focus more on scene layouts and propose in Sec. 3.1 a richer model in that aspect, including multiple lanes, crosswalks and sidewalks. Moreover, our base framework is able to infer model parameters with a single perspective image as input. A more recent work [14] proposes to infer a graph representation of the road, including lanes and lane markings, from partial segmentations of an image. Unlike our method that aims to handle complex road scenarios, [14] focuses only on straight roads. Máttyus *et al*. [17] propose an interesting parametric model of roads with the goal of augmenting existing map data with richer semantics. Again, this model only handles straight roads and requires input from both perspective and aerial images. Perhaps [23] is the closest work to ours. In contrast to it, we propose a fully-parametric model that is capable of reconstructing complex road layouts.

Learning from simulated data: Besides the scene model itself, one key contribution of our work is the training procedure that leverages simulated data, where we also utilize tools from domain adaptation [6, 30]. While most recent advances in this area focus on bridging domain gaps between synthetic and real RGB images [20, 21], we benefit from the semantic top-view representation within which our model is defined. This representation allows efficient modeling and sampling of a variety of road layouts, while avoiding the difficulty of photo-realistic renderings, to significantly reduce the domain gap between simulated and real data.

3. Our Framework

The goal of this work is to extract interpretable attributes of the layout of complex road scenes from a single camera. Sec. 3.1 presents our first contribution, a parameterized and rich model of road scenes describing attributes like the topology of the road, the number of lanes or distances to scene elements. The design of our scene model allows efficient sampling and, consequently, enables the generation of large-scale simulated data with accurate and complete annotations. At the same time, manual annotation of such scene attributes for real images is costly and, more importantly, even infeasible for some attributes, see Sec. 3.2. The second
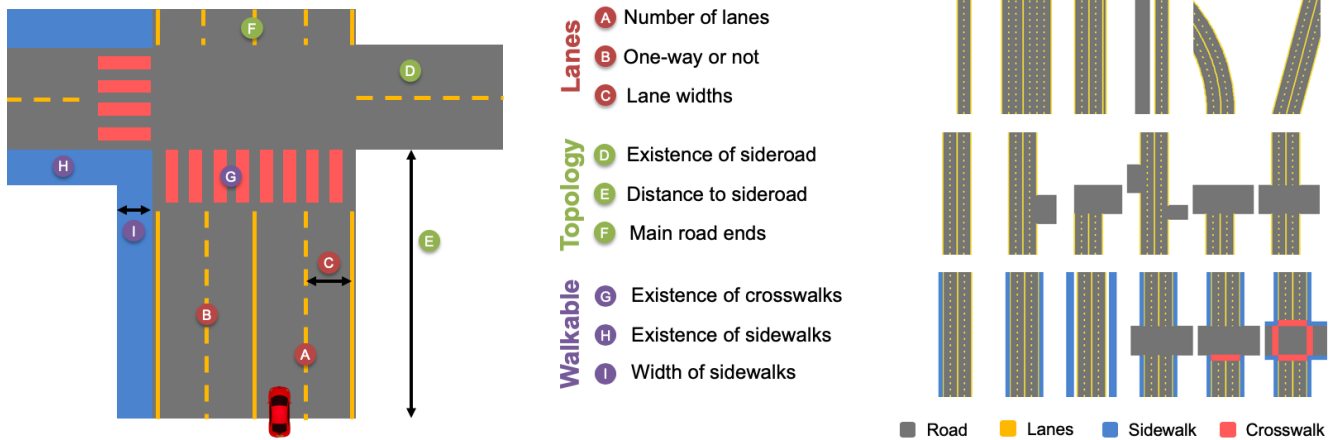
Figure 2: Our scene model consists of several parameters that capture a variety of complex driving scenes. (Left) We illustrate the model and highlight important parameters (A-I), which are grouped into three categories (middle): *Lanes*, to describe the layout of a single road; *Topology*, to model various road topologies; *Walkable*, describing scene elements for pedestrians. Our model is defined as a directed acyclic graph enabling efficient sampling and is represented in the top-view, making rendering easy. These properties turn our model into a simulator of semantic top-views. (Right) We show rendered examples for each of the above groups. A complete list of scene parameters and the corresponding graphical model is given in the supplementary.

contribution of our work, described in Sec. 3.3, is a deep learning framework that leverages training data from both domains, real and simulation, to infer the parameters of our proposed scene model. Finally, our third contribution is a conditional random field (CRF) that enforces coherence between related parameters of our scene model and encourages temporal smoothness for video inputs, see Sec. 3.4.

3.1. Scene Model

Our model describes road scenes in a semantic top-view representation and we assume the camera to be at the bottom center in every frame. This allows us to position all elements relative to the camera. On a higher level, we differentiate between the "main road", which is where the camera is, and eventual "side roads". All roads consist of at least one lane and intersections are a composition of multiple roads. Fig. 2 gives an overview of our proposed model.

Defining two side roads (one on the left and one on the right of the main road) along with distances to each one of them gives us the flexibility to model both 3-way and 4-way intersections. An additional attribute determines if the main road ends after the intersection, which yields T-intersections.

The main road is defined by a set of lanes, one- or two-way traffic, delimiters and sidewalks. We also define up to six lanes on the left and right side of the camera, which occupies the ego-lane. We allow different lane widths to model special lanes like turn- or bike-lanes. Next to the outer most lanes, optional delimiters of a certain width separate the road from the optional sidewalk. At intersections, we also model the existence of crosswalks at all four potential sides. For side roads, we only model their width. Our final

set of parameters Θ is grouped into different types and we count $M^{\mathrm{b}} = 14$ binary variables Θ_{b}, $M^{\mathrm{m}} = 2$ multi-class variables Θ_{m} and $M^{\mathrm{c}} = 22$ continuous variables Θ_{c}. The supplemental material contains a complete list of our model parameters. Note that the ability to work with a simple simulator means we can easily extend our scene model with further parameters and relationships.

3.2. Supervision from Real and Simulated Data

Inferring our model's parameters from an RGB image requires abundant training data. Seff and Xiao [24] leverage OpenStreetMaps [19] to gather ground truth for an RGB image. While this can be done automatically given the GPS coordinates, the set of attributes retrievable is limited and can be noisy. Instead, we leverage a combination of manual annotation and simulation for training.

Real data: Annotating real images with attributes corresponding to our defined parameters can be done efficiently only when suitable tools are used. This is particularly true for sequential data because many attributes stay constant over a long period of time. The supplemental material contains details on our annotation tool and process. We have collected a data set $D^{\mathrm{r}} = \{\mathbf{x}^{\mathrm{r}}, \Theta^{\mathrm{r}}\}_{i=1}^{N^{\mathrm{r}}}$ of N^{r} samples of semantic top-views $\mathbf{x}^{\mathrm{r}}$ and corresponding scene attributes Θ^{r}. The semantic top-views $\mathbf{x}^{\mathrm{r}} \in \mathbb{R}^{H \times W \times C}$, with spatial dimensions $H \times W$, contain C semantic categories ("road", "sidewalk", "lane boundaries" and "crosswalks") and are computed by applying the framework of [23]. However, several problems arise with real data. First, ground truth depth is required at a reasonable density for each RGB image to ask humans to reliably estimate distances to scene elements like
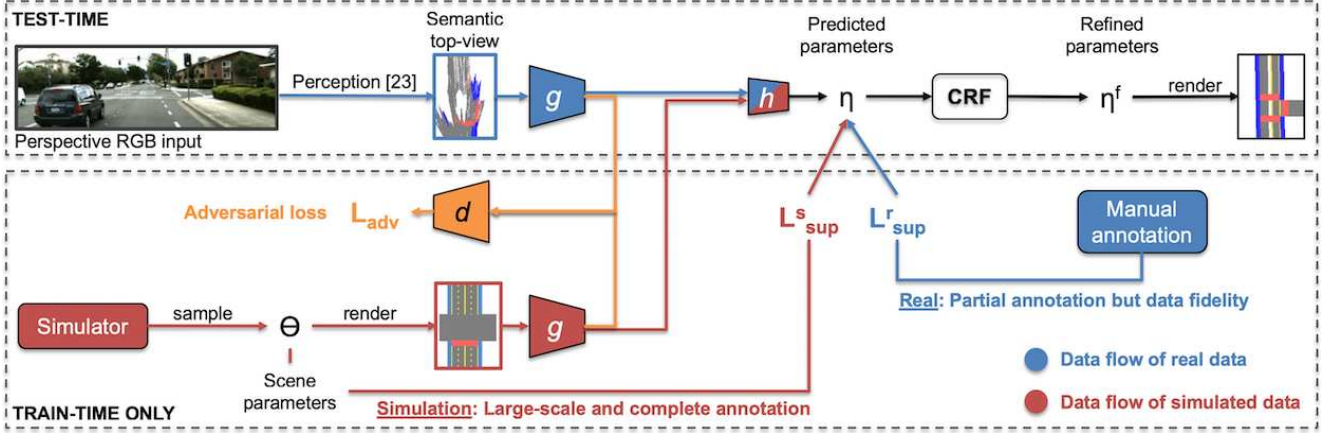
Figure 3: **Overview of our proposed framework:** At train-time, our framework makes use of both manual annotation for real data (blue) and automated annotation for simulated data (red), see Sec. 3.2. The feature extractors g convert semantic top views from either domain into a common representation which is input to h. An adversarial loss (orange) encourages a domain-agnostic output of g. At test-time, an RGB image in the perspective view is first transformed into a semantic top-view [23], which is then used by our proposed neural network (see Sec. 3.3), $h \circ g$, to infer our scene model (see Sec. 3.1). The graphical model defined in Sec. 3.4 ensures a coherent final output.

intersections or crosswalks. Second, there is always a limit on how much diverse data can be annotated cost-efficiently. Third, and most importantly, not all desired scene attributes are easy or even possible to annotate at a large-scale, even if depth information is available. For these reasons, we explore simulation as another source of supervision.

Simulated data: Our proposed scene model defined in Sec. 3.1 can act as a simulator to generate training data with complete and accurate annotation. First, by treating each attribute as a random variable with a certain hand-defined (conditional) probability distribution and relating them in a directed acyclic graph, we can use ancestral sampling [2] to efficiently sample a diverse set of scene parameters Θ^s. Second, we render the scene defined by the parameters Θ^s into a semantic top-view $\mathbf{x}^s$ with the same dimensions as $\mathbf{x}^r$. It is important to highlight that rendering is easy, compared to photo-realistic rendering of perspective RGB images [20, 21], because our model (i) works in the top-view where occlusion reasoning is not required and (ii) is defined in semantic space making illumination or photo-realism obsolete. We generate a data set $D^s = \{\mathbf{x}^s, \Theta^s\}_{i=1}^{N^s}$ of N^s simulated semantic top-views $\mathbf{x}^s$ and corresponding Θ^s. Fig. 2 (right) gives a few examples of rendered top-views.

3.3. Training and Inferring the Scene Model

We propose a deep learning framework that maps a semantic top-view $\mathbf{x}$ into the scene model parameters Θ. Fig. 3 provides a conceptual illustration. To leverage both sources of supervision (real and simulated data) during training, we define this mapping as

$$\Theta = f(\mathbf{x}) = (h \circ g)(\mathbf{x}) , \tag{1}$$

where $\circ$ defines a function composition and h and g are neural networks, with weights γ_h and γ_g respectively, that we want to train. The architecture of g is a 6-layer convolutional neural network (CNN) that converts a semantic top-view $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ into a 1-dimensional feature vector $f_{\mathbf{x}} \in \mathbb{R}^D$. Then, the function h is defined as a multi-layer perceptron (MLP) predicting the scene attributes Θ given $f_{\mathbf{x}}$. Specifically, h 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.

Our objective is that $h \circ g$ works well on real data, while we want to leverage the rich and large set of annotations from simulated data during training. The intuition behind our design is to have a unified g that maps semantic top-views $\mathbf{x}$ of different domains into a common feature representation, usable by a domain-agnostic classifier h. To realize this intuition, we define supervised loss functions on both real and simulated data and leverage domain adaptation techniques to minimize the domain gap between the output of g given top-views from different domains.

Loss functions on scene attribute annotation: Given data sets D^r and D^s of real and simulated data, we define

$$\mathcal{L}_{\text{sup}} = \lambda^r \cdot \mathcal{L}_{\text{sup}}^r + \lambda^s \cdot \mathcal{L}_{\text{sup}}^s \tag{2}$$

as supervised loss. The scalars λ^r and λ^s weigh the importance between real and simulated data and

$$
\begin{aligned}
\mathcal{L}_{\text{sup}}^{\{r,s\}} = \sum_{i=1}^{N^{\{r,s\}}} &\text{BCE}(\Theta_{b,i}^{\{r,s\}}, \eta_{b,i}^{\{r,s\}}) \\
&+ \text{CE}(\Theta_{m,i}^{\{r,s\}}, \eta_{m,i}^{\{r,s\}}) \\
&+ \ell_1(\Theta_{c,i}^{\{r,s\}}, \eta_{c,i}^{\{r,s\}}) ,
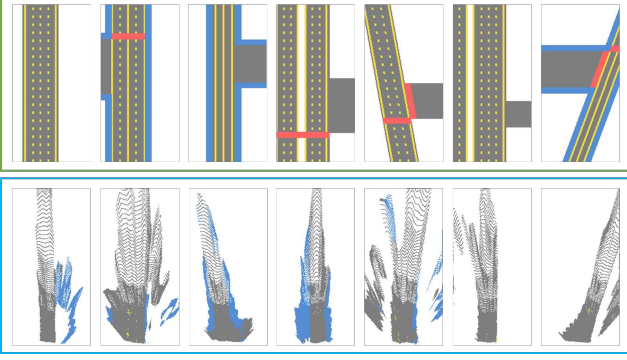\end{aligned}
\tag{3}
$$

Figure 4: Unpaired examples of simulated semantic top-views (top) and real ones from [23] (bottom).

where (B)CE is the (binary) cross-entropy loss and $\{\Theta, \eta\}_{\cdot,i}$ denotes the i-th sample in the data set. For regression, we discretize continuous variables into K bins by convolving a dirac delta function centered at Θ_c with a Gaussian of fixed variance, which enables easier multi-modal predictions and is useful for the graphical model defined in Sec. 3.4. We ignore scene attributes without manual annotation for $\mathcal{L}_{sup}^r$.

Bridging the domain gap: Since our goal is to leverage simulated data during the training process, our network design needs to account for the inherent domain gap. We thus define separate feature extraction networks g_r and g_s with shared weights γ_g that take as input semantic top-views from either domain, *i.e.*, $\mathbf{x}^r$ or $\mathbf{x}^s$, and compute respective features $f_{\mathbf{x}^r}$ and $f_{\mathbf{x}^s}$. We then explicitly encourage a domain-agnostic feature representation by employing an adversarial loss function $\mathcal{L}_{adv}$ [6]. We use an MLP $d(f_{\mathbf{x}})$ with parameters γ_d as discriminator, that takes the feature representations from either domain, *i.e.*, $f_{\mathbf{x}^r}$ or $f_{\mathbf{x}^s}$, as input and makes a binary prediction into "real" or "fake". As in standard generative adversarial networks, d has the goal to discriminate between the two domains, while the rest of the model aims to confuse the discriminator by providing inputs $f_{\mathbf{x}^{r,s}}$ indistinguishable in the underlying distribution, *i.e.*, a domain-agnostic representation of the semantic top-view maps $\mathbf{x}^{r,s}$. Fig. 4 shows *unpaired* examples of simulated and real top-views to illustrate the domain gap. Note that we chose domain adaptation on the feature-level over our initial attempt on the pixel-level with a modified version of [35] due to a simpler design and higher accuracy. However, we refer to the supplementary for a discussion of our pixel-level approach, which provides insights into the role of domain adaptation and further visualizations on overcoming the domain gap.

Optimization: We use ADAM [13] to estimate the parameters of our neural network model by solving:

$$\max_{\gamma_d} \min_{\gamma_g, \gamma_h} \mathcal{L}_{sup} + \lambda^{adv} \mathcal{L}_{adv}. \qquad (4)$$

Figure 3 provides an overview of our framework.

3.4. CRF for Coherent Scene Understanding

We now introduce our graphical model for predicting consistent layouts of road scenes. We first present our CRF for single frames and then extend it to the temporal domain.

Single image CRF: Let us first denote the elements of scene attributes and corresponding predictions as $\Theta[\cdot]$ and $\eta[\cdot]$, where we use indices $i \in \{1, ..., M^b\}$, $p \in \{1, ..., M^m\}$ and $m \in \{1, ..., M^c\}$ for binary, multi-class and continuous variables, respectively. We then formulate scene understanding as the energy minimization problem

$$\begin{aligned} E(\Theta|\mathbf{x}) = & E_b(\Theta_b) + E_m(\Theta_m) + E_c(\Theta_c) \\ & + E_s(\Theta_b, \Theta_m) + E_q(\Theta_b, \Theta_c) \qquad (5) \\ & + E_h(\Theta_b, \Theta_m, \Theta_c) , \end{aligned}$$

where E_* denotes energy potentials for the associated scene attribute variables (Θ_b, Θ_m and Θ_c). We will describe the details for each of those potentials in the following.

For binary variables Θ_b, our potential function E_b consists of two terms,

$$E_b(\Theta_b) = \sum_i \phi_b(\Theta_b[i]) + \sum_{i \neq j} \psi_b(\Theta_b[i], \Theta_b[j]) . \qquad (6)$$

The unary term $\phi_b(\cdot)$ specifies the cost of assigning a label to Θ_b^i and is defined as $-\log P_b(\Theta_b[i])$, where $P_b(\Theta_b[i]) = \eta_b[i]$ is the probabilistic output of our neural network h. The pairwise term $\psi_b(\cdot, \cdot)$ defines the cost of assigning $\Theta_b[i]$ and $\Theta_b[j]$ to i-th and j-th variable as $\psi_b(\Theta_b[i], \Theta_b[j]) = -\log M_b(\Theta_b[i], \Theta_b[j])$, where M_b is the co-occurrence matrix and $M_b(\Theta_b[i], \Theta_b[j])$ is the corresponding probability. For multi-class variables, our potential is defined as $E_m(\Theta_m) = \sum_p \phi_m(\Theta_m[p])$, where $\phi_m(\cdot) = -\log P_m(\cdot)$ and $P_m(\Theta_m[p]) = \eta_m[p]$. Similarly, we define the potential for continuous variables as $E_c(\Theta_c) = \sum_m \phi_c(\Theta_c[m])$ with $\phi_c(\Theta_c[m])$ being the negative log-likelihood of $\eta_c[m]$.

For a coherent prediction, we further introduce the potentials E_s, E_q and E_h to model correlations among scene attributes. E_s and E_q enforce hard constraints between certain binary variables and multi-class or continuous variables, respectively. They convey the idea that, for instance, the number of lanes of a side-road is consistent with the actual existence of that side-road. We denote the set of pre-defined pairs between Θ_b and Θ_m as $\mathcal{S} = \{(i, p)\}$ and between Θ_b and Θ_c as $\mathcal{Q} = \{(i, m)\}$. Potential E_s is then defined as

$$E_s(\Theta_b, \Theta_m) = \sum_{(i,p) \in \mathcal{S}} \infty \times \mathbb{1}[\Theta_b[i] \neq \Theta_m[p]] , \qquad (7)$$

where $\mathbb{1}[*]$ is the indicator function. Potential E_q is defined likewise but using the set $\mathcal{Q}$ and variables Θ_c. In both cases, we give a high penalty to scenarios where two types of predictions are inconsistent.

Finally, the potential E_h of our energy defined in Eq. (5) models higher-order relations between Θ_b, Θ_m and Θ_c. The potential takes the form

$$E_h(\Theta_b, \Theta_m, \Theta_c) = \sum_{c \in \mathcal{C}} \infty \times f_c(\Theta_b[i], \Theta_m[p], \Theta_c[m]) \ , \quad (8)$$

where $c = (i, p, m)$ and $f_c(\cdot, \cdot, \cdot)$ is a table where conflicting predictions are set to 1. The supplementary gives a complete definition of $\mathcal{C}$ which contains the relations between scene attributes and the constraints we enforce on them.

Temporal CRF: Given videos as input, we propose to extend our CRF to encourage temporally consistent and meaningful outputs. We extend the energy function from Eq. (5) by two terms that enforce temporal consistency of binary and multi-class variables and smoothness for continuous variables. Due to space limitations, we refer to the supplementary for details of our formulation.

Learning and inference on CRF: We use QPBO [22] for inference in both CRF models. Since ground truth is not available for all frames, we do not introduce per-potential weights. However, our CRF is amenable to piece-wise [28] or joint learning [5, 34] if ground-truth is provided.

4. Experiments

To evaluate the quality of our scene understanding approach we conduct several experiments and analyze the importance of different aspects of our model. Since we do have manually-annotated ground truth, we can quantify our results and compare with several baselines that demonstrate the impact of two key contributions: the use of top-view maps and simulated data for training. We also put an emphasis on qualitative results in this work for two reasons: First, not all attributes of our model are actually contained in the manually-annotated ground truth and can thus not be quantified but only qualitatively verified. Second, there is obviously no prior art showing results on this novel set of ground truth data, which makes the analysis of qualitative results even more important.

Datasets: Since our focus is on driving scenes and our approach requires semantic segmentation and depth annotation, we choose to work with the KITTI [8] and the newly released NuScenes [18][2] data sets. Although both data sets provide laser-scanned data for depth ground truth, note that depth supervision can also come from stereo images [9]. Also, since NuScenes [18] does not provide semantic segmentation, we reuse the segmentation model from KITTI. For both data sets, we manually annotate a subset of the images with our scene attributes. Annotators see the RGB image as well as the depth ground truth and provide labels for 22 attributes

of our model. We refer to the supplementary for details on the annotation process. In total, we acquired around 17000 annotations for KITTI [8] and 3000 for NuScenes [18].

Evaluation metrics: Since the output space of our prediction is complex and consists of a mixture of discrete and continuous variables, which require different handling, we use multiple different metrics for evaluation.

For binary variables (like the existence of side roads) and for multi-class variables (like the number of lanes), we measure accuracy as Accu.-Bi $= \frac{1}{14} \sum_{k=1}^{14} [p_k = \Theta_{bk}]$ and Accu.-Mc $= \frac{1}{2} \sum_{k=1}^{2} [p_k = \Theta_{mk}]$. For regression variables we use the mean square error (MSE).

Besides these standard metrics, we also propose another metric that combines all predicted variables and outputs into a single number. We take the predicted parameters and render the scene accordingly. For the corresponding image, we take the ground truth parameters (augmented with predicted values for variables without ground truth annotation) and render the scene, which assigns each pixel a semantic category. For evaluation, we can now use Intersection-over-Union (IoU), a standard measure in semantic segmentation. While being a very challenging metric in this setup, it implicitly weighs the attributes by their impact on the area of the top-view. For instance, predicting the number of lanes incorrectly by one has a bigger impact than getting the distance to a sideroad wrong by one meter.

4.1. Single Image Evaluation

Our main experiments are conducted with a single image as input. In the next section, we separately evaluate the impact of temporal modeling as described in Sec. 3.4.

Baselines: Since we propose a scene model of roads with new attributes and corresponding ground truth annotation, there exist no previously reported numbers. We thus choose appropriate baselines that are either variations of our model or relevant prior works extended to our scene model:

- **Manual-GT-RGB (M-RGB):** A classification CNN (ResNet-101 [12]) trained on the manually-annotated ground truth. Seff and Xiao [24] have the same setup except that we use a network with more parameters and train for all attributes simultaneously in a multi-task setup.
- **Manual-GT-RGB+Depth (M-RGB+D):** Same as M-RGB but with the additional task of monocular depth prediction (as in our perception model). The intuition is that this additional supervision aids predicting certain scene attributes, *e.g.*, distances to side roads, and renders a more fair comparison point to our model.
- **Manual-GT-BEV (M-BEV):** Instead of using the perspective RGB image as input, this baseline uses the output of [23], which is a semantic map in the top-view, also referred to as bird's eye view (BEV). We train the function f with the manually annotated ground truth. Thus, M-BEV can be seen as an extension of [23] to our scene model.
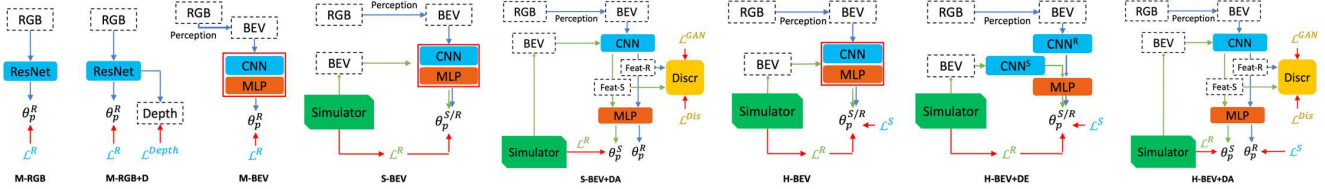
Figure 5: Illustrations of all the models we compare in the quantitative evaluation in Tab. 1.

Method	KITTI [8]				NuScenes [18]			
	Accu.-Bi. ↑	Accu.-Mc. ↑	MSE ↓	IOU ↑	Accu.-Bi. ↑	Accu.-Mc. ↑	MSE ↓	IOU ↑
M-RGB [24]	.811	.778	.230	.317	.846	.604	.080	.316
M-RGB [24]+D	.799	.798	.146	.342	.899	.634	.021	.335
M-BEV [23]	.820	.777	.141	.345	.852	.601	.022	.269
M-BEV [23] +GM	.831	.792	.136	.350	.852	.601	.036	.338
S-BEV	.694	.371	.249	.239	.790	.366	.162	.155
S-BEV+DA	.818	.677	.222	.314	.753	.568	.103	.171
S-BEV+DA+GM	.847	.683	.230	.320	.723	.568	.081	.160
H-BEV	.816	.756	.152	.342	.783	.569	.039	.345
H-BEV+DE	.830	.776	.158	.381	.854	.626	.042	.423
H-BEV+DA	.845	.792	.108	.398	.856	.545	.028	.346
H-BEV+DA+GM	.849	.805	.098	.371	.855	.626	.033	.450

Table 1: Main results on road scene layout estimation on both data sets KITTI [8] and NuScenes [18].

- **Simulation-BEV (S-BEV):** This baseline uses the same architecture as M-BEV but is trained only in simulation.
- **Simulation-BEV+DomainAdapt (S-BEV+DA):** Same as S-BEV, but with additional domain adaptation loss as proposed in our model.

We denote our approach proposed in Sec. 3, according to the nomenclature above, as **Hybrid-BEV+DomainAdapt (H-BEV+DA)** and further explore two variants of it. First, **H-BEV** does not employ the discriminator d but still trains from both domains. Second, **H-BEV+DE** also avoids the discriminator but uses a separate set of weights γ_g^r and γ_g^s for the feature extraction network g. The intuition is that the supervised losses from both domains and the separate domain-specific encoding (thus, "+DE") already provide enough capacity and information to the model to find a domain-agnostic representation of the data. Please refer to Fig. 5 for an overview of the different models we compare. For the best models among each group (M-, S- and H-), we report numbers with the graphical model (+GM).

Quantitative results: Tab. 1 summarizes our main results for both data sets and we can draw several conclusions. First, when comparing the groups of methods by supervision type, *i.e.*, manual (M), simulation (S) and hybrid (H), we can clearly observe the benefit of hybrid methods leveraging both domains. Second, within the group of manual annotation, we can see that adding depth supervision to the approach of [24] significantly improves results, particularly for continuous variables. Predicting the scene attributes directly from the top-view representation of [23] is slightly better than M-

Method	KITTI [8]		NuScenes [18]	
	seman.↓	temp.↓	seman.↓	temp.↓
S-BEV+DA	2.82	5.32	1.08	2.09
M-BEV [23]	2.65	3.99	1.09	1.27
H-BEV+DA	5.59	6.01	1.08	1.05
+GM	1.77	1.93	0.11	0.42

Table 2: Main results on consistency measurements.

RGB+D on KITTI and worse on NuScenes, but has the crucial advantage that augmentation with simulated data in the top-view becomes possible, as illustrated with all hybrid variants. Third, within the group of simulated data, using domain adaptation techniques (S-BEV+DA) has a significant benefit. We want to highlight the competitive overall results of S-BEV+DA, which is an unsupervised domain adaptation approach requiring no manual annotation. Forth, also for hybrid methods, explicitly addressing the domain gap (H-BEV+DE and H-BEV+DA) enables higher accuracy. Finally, all models improve with our graphical model put on top.

Qualitative results: We show several qualitative results in Fig. 6 and Fig. 7 and again highlight their importance to demonstrate the practicality of our approach qualitatively. We can see from the examples that our model successfully describes a diverse set of road scenes.

4.2. Evaluating consistency of our model

We now analyze the impact of the graphical model on the consistency of our predictions, for which we define the following metrics:
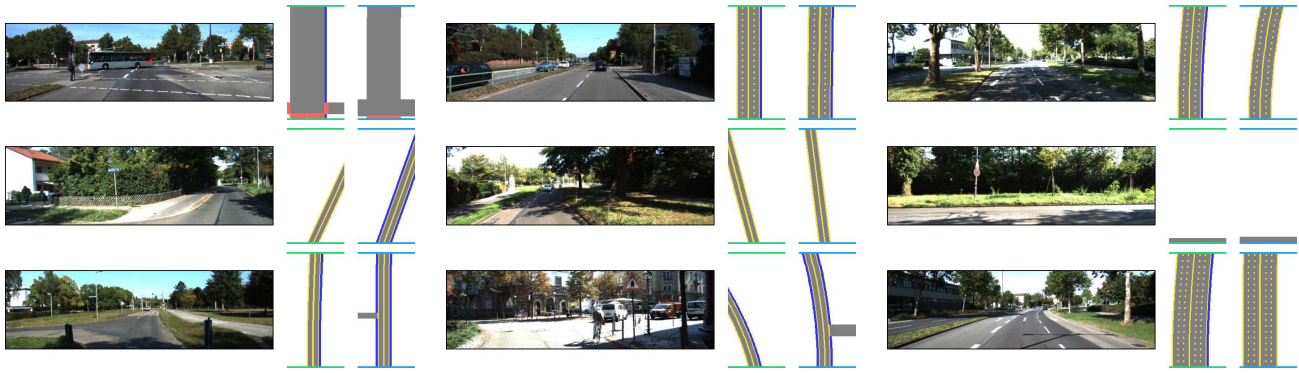
Figure 6: Qualitative results of H-BEV+DA+GM on individual frames from KITTI. Each example shows perspective RGB, ground truth and predicted semantic top-view, respectively. Our representation is rich enough to cover various road layouts and handles complex scenarios, *e.g.*, rotation, existence of crosswalks, sidewalks, side-roads and curved roads.

Figure 7: Qualitative results comparing H-BEV+DA and H-BEV+DA+GM in consecutive frames of two example sequences of the KITTI validation set. In each column, we have visualized the perspective RGB image, prediction from H-BEV+DA and that of H-BEV+DA+GM from left to right. Each row shows a sequence of three frames. We can observe more consistent predictions, *e.g.*, width of side-road and delimiter width, with the help of the temporal CRF.

- *Semantic consistency:* we measure the conflicts in attribute predictions w.r.t. their semantic meanings. Specifically, we count a conflict if predicted attributes are not feasible in our scene model. The average number of conflicts is reported as our semantic consistency measurement.

- *Temporal consistency:* for each attribute prediction among a video sequence, we measure the number of changes in the prediction. We report the average number of prediction changes as the temporal consistency. The lower the number is, the more stable prediction we would obtain. Note that consistency itself cannot replace the accuracy since a prediction can also be consistently wrong.

As for the temporal consistency, we visualize qualitative results of consecutive frames in two validation sequences from KITTI in Fig. 7. The graphical model successfully enforces temporal smoothness, especially for number of lanes, delimiter width and the width of side-roads.

Finally, we show in Tab. 2 quantitative results for the temporal consistency metrics defined above on both KITTI and NuScenes data sets. We compare representative models from each group of different forms of supervision (M-, S- and H-) with the output of the graphical model applied on H-BEV+DA. We can clearly observe a significant improvement in consistency for both data sets. Together with the superior results in Tab. 1, this clearly demonstrates the benefits of the

proposed graphical model for our application.

5. Conclusion

In this work, we present a scene understanding framework for complex road scenarios. Our key contributions are: (1) A parameterized and interpretable model of the scene that is defined in the top-view and enables efficient sampling of diverse scenes. The semantic top-view representation makes rendering easy (compared to photo-realistic RGB images in perspective view), which enables the generation of large-scale simulated data. (2) A neural network design and corresponding training scheme to leverage both simulated as well as manually-annotated real data. (3) A graphical model that ensures coherent predictions for a single frame input and temporally smooth outputs for a video input. Our proposed hybrid model (using both sources of data) outperforms its counterparts that use only one source of supervision in an empirical evaluation. This confirms the benefits of the top-view representation, enabling simple generation of large-scale simulated data and consequently our hybrid training.

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] Christopher M. Bishop. *Pattern Recogntion and Machine Learning*. Springer, 2007.

[3] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. In-Place Activated BatchNorm for Memory-Optimized Training of DNNs. In *CVPR*, 2018.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *ECCV*, 2018.

[5] Justin Domke. Learning graphical model parameters with approximate marginal inference. *PAMI*, 35(10):2454–2467, 2013.

[6] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain adversarial training of neural networks. *JMLR*, 2016.

[7] Andreas Geiger, Martin Lauer, Christian Wojek, Christoph Stiller, and Raquel Urtasun. 3D Traffic Scene Understanding from Movable Platforms. *PAMI*, 2014.

[8] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets Robotics: The KITTI Dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[9] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised Monocular Depth Estimation with Left-Right Consistency. In *CVPR*, 2017.

[10] Ruiqi Guo and Derek Hoiem. Beyond the line of sight: labeling the underlying surfaces. In *ECCV*, 2012.

[11] Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive Mapping and Planning for Visual Navigation. In *CVPR*, 2017.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015.

[14] 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.

[15] Iro Laina, Vasileios Belagiannis Christian Rupprecht, Federico Tombari, and Nassir Navab. Deeper Depth Prediction with Fully Convolutional Residual Networks. In *3DV*, 2016.

[16] Chenxi Liu, Alexander G. Schwing, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Rent3D: Floor-Plan Priors for Monocular Layout Estimation. In *CVPR*, 2015.

[17] Gellért Máttyus, Shenlong Wang, Sanja Fidler, and Raquel Urtasun. HD Maps: Fine-grained Road Segmentation by Parsing Ground and Aerial Images. In *CVPR*, 2016.

[18] NuTonomy. The NuScenes data set. `https://www.nuscenes.org`, 2018.

[19] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`, 2017.

[20] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for Benchmarks. In *ICCV*, pages 2232–2241. IEEE, 2017.

[21] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for Data: Ground Truth from Computer Games. In *ECCV*, pages 102–118. Springer, 2016.

[22] Carsten Rother, Vladimir Kolmogorov, Victor Lempitsky, and Martin Szummer. Optimizing binary MRFs via extended roof duality. In *CVPR*. IEEE, 2007.

[23] Samuel Schulter, Menghua Zhai, Nathan Jacobs, and Manmohan Chandraker. Learning to Look around Objects for Top-View Representations of Outdoor Scenes. In *ECCV*, 2018.

[24] Ari Seff and Jianxiong Xiao. Learning from Maps: Visual Common Sense for Autonomous Driving. *arXiv:1611.08583*, 2016.

[25] Sunando Sengupta, Paul Sturgess, Ĺubor Ladický, and Philip H. S. Torr. Automatic Dense Visual Semantic Mapping from Street-Level Imagery. In *IROS*, 2012.

[26] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic Scene Completion from a Single Depth Image. In *CVPR*, 2017.

[27] 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.

[28] Charles Sutton and Andrew McCallum. Piecewise training for undirected models. In *UAI*. AUAI Press, 2005.

[29] Joseph Tighe, Marc Niethammer, and Svetlana Lazebnik. Scene Parsing with Object Instances and Occlusion Ordering. In *CVPR*, June 2014.

[30] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to Adapt Structured Output Space for Semantic Segmentation. In *CVPR*, 2018.

[31] Shubham Tulsiani, Richard Tucker, and Noah Snavely. Layer-structured 3D Scene Inference via View Synthesis. In *ECCV*, 2018.

[32] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. Structured Attention Guided Convolutional Neural Fields for Monocular Depth Estimation. In *CVPR*, 2018.

[33] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid Scene Parsing Network. In *CVPR*, 2017.

[34] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015.

[35] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.