Projecting Your View Attentively: Monocular Road Scene Layout Estimation via Cross-view Transformation

Weixiang Yang[1],Qi Li[1],Wenxi Liu[1*],Yuanlong Yu[1*],Yuexin Ma[2,3],Shengfeng He[4],Jia Pan[5]
[1]*College of Mathematics and Computer Science, Fuzhou University*[*]
[2]*ShanghaiTech University* [3]*Shanghai Engineering Research Center of Intelligent Vision and Imaging*
[4]*School of Computer Science and Engineering, South China University of Technology*
[5]*Department of Computer Science, The University of Hong Kong*

Abstract

HD map reconstruction is crucial for autonomous driving. LiDAR-based methods are limited due to the deployed expensive sensors and time-consuming computation. Camera-based methods usually need to separately perform road segmentation and view transformation, which often causes distortion and the absence of content. To push the limits of the technology, we present a novel framework that enables reconstructing a local map formed by road layout and vehicle occupancy in the bird's-eye view given a front-view monocular image only. In particular, we propose a cross-view transformation module, which takes the constraint of cycle consistency between views into account and makes full use of their correlation to strengthen the view transformation and scene understanding. Considering the relationship between vehicles and roads, we also design a context-aware discriminator to further refine the results. Experiments on public benchmarks show that our method achieves the state-of-the-art performance in the tasks of road layout estimation and vehicle occupancy estimation. Especially for the latter task, our model outperforms all competitors by a large margin. Furthermore, our model runs at 35 FPS on a single GPU, which is efficient and applicable for real-time panorama HD map reconstruction.

1. Introduction

With the rapid progress of autonomous driving technologies, many recent efforts have been spent on the related research topics, e.g., scene layout estimation [11, 22, 24, 36, 38, 42, 48], 3D object detection [5, 6, 21, 37, 40, 41], vehicle behavior prediction [15, 19, 26, 30], and lane detection [16, 33, 57], etc.

Among these tasks, high-definition map (HD map) reconstruction is fundamental and critical for perception, prediction, and planning of autonomous driving. Its major issues are concerned with the estimation of a local map including the road layout as well as the occupancies of nearby vehicles in the 3D world. Existing techniques rely on expensive sensors like LiDAR and require time-consuming computation for cloud point data. Besides, the camera-based techniques usually need to separately perform road segmentation and view transformation, which thus causes distortion and the absence of content.

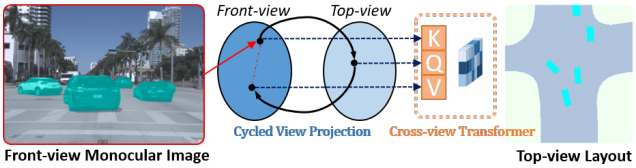


Figure 1. Given a frontal view monocular image, we propose to leverage a cycle structure that bridges the features of frontal view and top view in their respective domains, as well as a cross-view transformer that correlates views attentively in order to facilitate the road layout estimation.

To push the limits of the technology, our work aims to address this realistic yet challenging problem of estimating the road layout and vehicle occupancy in top view or bird's-eye view (BEV), given a single monocular front-view image (see Fig. 1). However, due to the large view gap and severe view deformation, understanding and estimating the top-view scene layout from the front-view image is an extremely difficult problem even for a human observer. Particularly, the same scene has significantly different appearances in the images of bird's-eye view and frontal view. Thus, parsing and projecting the road scenes of frontal view to top view require the ability of fully exploiting the information of the frontal view image and innate reasoning the unseen regions.

Traditional methods (e.g. [23, 45]) focus on investigating the perspective transformation by estimating the camera parameters and performing image coordinate transformation, but but gaps in the resulting BEV feature maps caused by geometric warping lead to poor results. Recent deep learning based approaches [35, 56] mainly rely on the hallucination capability of deep Convolutional Neural Networks

[*]Wenxi Liu and Yuanlong Yu are the corresponding authors.

(CNNs) to infer the unseen regions between views. In general, instead of modeling the correlation between views, these methods directly leverage CNNs to learn the view projection models in a supervised manner. These models require deep network structures to propagate and transform the features of frontal view through multiple layers to spatially align with the top-view layout. However, due to the locally confined receptive fields of convolutional layers, it causes the difficulty of fitting a view projection model and identifying the vehicles of small scales. Moreover, road layout provides the crucial context information to infer the position and orientation of vehicles, e.g., vehicles parked alongside the road. Yet, the prior road scene parsing methods usually ignore the spatial relationship between vehicles and roads.

To address the aforementioned concerns, we derive a novel GAN-based framework to estimate the road layout and vehicle occupancies from top view, given a single monocular front-view image. To handle the large discrepancy between views, we present a cross-view transformation module in the generator network, which is composed of two sub-modules: *Cycled View Projection* (CVP) module bridges the view features in their respective domains and *Cross-View Transformer* (CVT) correlates the views, as shown in Fig. 1. Specifically, the CVP utilizes a multi-layer perceptron (MLP) to project views, which overtakes the standard information flow passing through convolutional layers, and involves the constraint of cycle consistency to retain the features relevant for view projection. In other word, transforming frontal views to top views requires a global spatial transformation over the visual features. Yet, standard CNN layers only allow local computation over feature maps, which thus takes several layers to obtain a sufficiently large receptive field. On the other hand, fully connected layers can better facilitate the cross-view transformation. Then, CVT explicitly correlates the features of the views before and after projection obtained from CVP, which can significantly enhance the features after view projection. In particular, we involve a feature selection scheme in CVT which leverages the associations of both views to extract the most relevant information. Furthermore, to exploit the spatial relationship between vehicles and roads, we present a context-aware discriminator that evaluates not only the estimated masks of vehicles but also their correlation.

In experimental results, we show that our cross-view transformation module and the context-aware discriminator can elevate the performance of road layout and vehicle occupancy estimation. For both tasks, we compare our model against the state-of-the-art methods on public benchmarks and demonstrate that our model is superior to all the other methods. It is worth noting that, for the estimation of vehicle occupancies, our model achieves a significant advantage over the other comparison methods by at least 28.5% in the *KITTI 3D Object* dataset and by at least 48.8% in the *Argoverse* dataset. We also show that our framework is able to process 1024×1024 images in 35 FPS using a single GPU, and it is applicable for real-time reconstruction of panorama HD map. The contributions of our paper are summarized as:

- We propose a novel framework that reconstructs a local map formed by top-view road scene layout and vehicle occupancy using a single monocular front-view image only. In particular, we propose a cross-view transformation module which leverages the cycle consistency between views and their correlation to strengthen the view transformation.
- We also propose a context-aware discriminator that considers the spatial relationship between vehicles and roads in the task of estimating vehicle occupancies.
- On public benchmarks, it is demonstrated that our model achieves the state-of-the-art performance for the tasks of road layout and vehicle occupancy estimation.

2. Related Work

In this section, we survey the related literature on road layout estimation, vehicle detection, and street view synthesis on top view representation. We also introduce the recent progress of transformers on vision tasks.

BEV-based Road layout estimation and vehicle detection. Most road scene parsing works focus on semantic segmentation [8, 9, 44, 50, 51], while there are a few attempts that derive top view representation for road layout [11, 13, 25, 28, 38, 39, 48, 52]. Amongst these methods, Schulter et al. [38] propose to estimate an occlusion-reasoned road layout on top view from a single color image by depth estimation and semantic segmentation. [25] proposes a variational autoencoder (VAE) model to predict road layout from a given image, yet without attempting to reason about the unseen layout from observation. Pan et al. [32] present a cross-view semantic segmentation by transforming and fusing the observation from multiple cameras. [34,36] directly transform features from images to 3D space and finally to bird's-eye-view (BEV) grids. On the other hand, many monocular image-based 3D vehicle detection techniques have been developed (e.g., [5, 21, 29]). Several methods handle this problem by mapping the monocular image to the top view. For instance, [37] proposes to map a monocular image to the top view representation, and treats 3D object detection as a task of 2D segmentation. BirdGAN [41] also leverages adversarial learning for mapping images to bird's-eye view. As another related work, [47] does not focus on explicit scene layout estimation, focusing instead more on the motion planning side. Most related to our work, [27] presents a unified model to

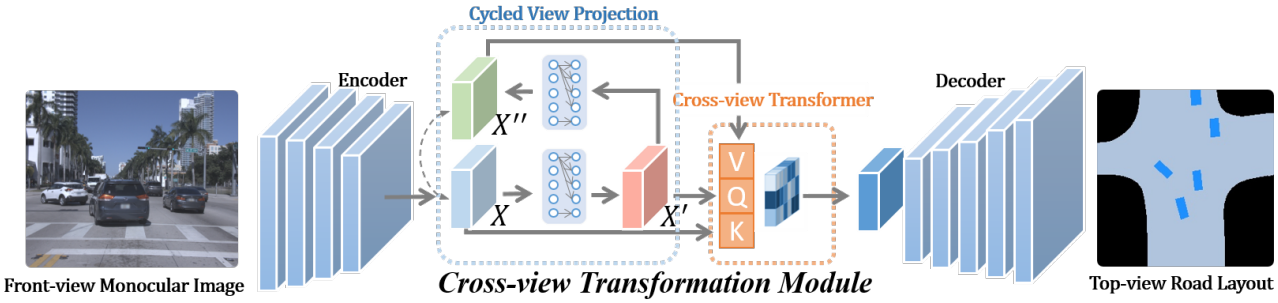


Figure 2. As illustrated, our network aims to transform the front-view monocular image to the top-view road layout. As the main component, our proposed *cross-view transformation module* consists of the cycled view projection (CVP) and the cross-view transformer (CVT), which projects the features from the front-view domain, X, to the top-view domain, X'. In CVP, it first utilizes a MLP-based cycle structure to retain the confident features for view projection in X'', and then CVT correlates the features of both views to attentively enhance X'.

tackle the task of road layout (static scene) and traffic participant (dynamic scene) estimations from a single image. Differently, we propose an approach to explicitly model the large view projection and exploits the context information for producing high-quality results.

View transformation and synthesis. Traditional methods (e.g. [17, 23, 45]) have been proposed to handle the perspective transformation in traffic scenes. With the progress of deep learning based methods, [56] proposes a pioneering work to generate the bird's-eye view based on the driver's view. They treat cross-view synthesis as an image translation task and adopt a GAN-based framework to accomplish it. Due to the difficulty in collecting annotation for real data, their model is trained from the data of video games. [1] focuses exclusively on warping camera images to BEV images without performing any downstream tasks such as object detection. Recent attempts [35, 43] on view synthesis aim to convert aerial images to street view images, or vice versa. Compared with these works, our purpose is quite different, which requires not only the implicit view projection from frontal view to top view, but also the estimation of road layout and vehicle occupancies under a unified framework.

Transformer for vision tasks. With recent success of the transformer [46], its ability of explicitly modeling pairwise interactions for elements in a sequence has been leveraged in many vision tasks [3, 10, 49, 53, 55]. Unlike these transformer-based models, our proposed cross-view transformer attempts to establish the correlation between the features of views. In addition, we incorporate a feature selection scheme along with the non-local cross-view correlation scheme, which significantly enhances the representativeness of the features.

3. Our Proposed Method

3.1. Network Overview

The goal of our work is to estimate the road scene layout and vehicle occupancies on the bird's-eye view in the form of semantic masks, given a monocular frontal view image.

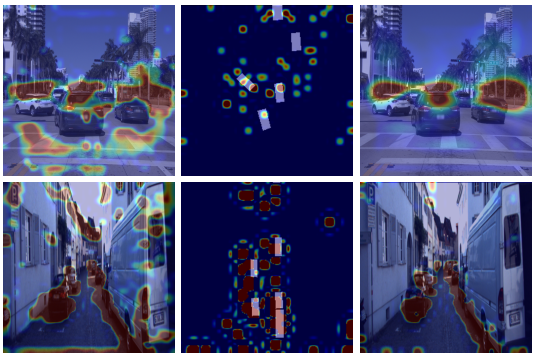
Our network architecture is based on a GAN-based framework. In specific, as shown in Fig. 2, the generator is an encoder-decoder architecture, in which the input frontal view image I is first passed through the encoder that adopts ResNet [14] as the backbone network to extract visual features, then our proposed cross-view transformation module that enhances the features for view projection, and finally the decoder to produce the top-view masks $\hat{M}$. On the other hand, we propose a context-aware discriminator (see Fig. 5) that discriminates against the masks of vehicles by taking the road context into account.In the following subsections, we will elaborate the details of our cross-view transformation module and the context-aware discriminator.

3.2. Cross-view Transformation Module

Due to the large gap between frontal views and top views, there exists a large amount of missing image content during view projection, so the traditional view projection techniques lead to defective results. To this end, the hallucination ability of CNN-based methods has been exploited to address the problem, but the patch-level correlation of both views is not trivial to model within deep networks.

In order to strengthen the view correlation while exploiting the capability of deep networks, we introduce a cross-view transformation module into the generator of GAN-based framework, which enhances the extracted visual features for projecting frontal view to top view. The structure of our proposed cross-view transformation module is shown in Fig. 2, which is composed of two parts: cycled view projection and cross-view transformer.

Cycled View Projection (CVP). Since the features of frontal views are not spatially aligned with the ones of top views due to their large gap, following the practice of [31], we deploy the MLP structure consisting of two fully-connected layers to project the features of frontal view to top view, which can overtake the standard information flow of stacking convolution layers. As shown in Fig. 2, X and X' represent the feature maps before and after view projection, respectively. Hence, the holistic view projection can



X X' X''

Figure 3. Visualization of the features at front view and top view by aligning them with the images of the corresponding views.

be achieved by: $X' = \mathcal{F}_{MLP}(X)$, where X refers to the features extracted from the ResNet backbone.

However, such a simple view projection structure cannot guarantee the information of frontal views to be effectively delivered. Here, we introduce a cycled self-supervision scheme to consolidate the view projection, which projects the top-view features back to the domain of frontal views. As illustrated in Fig. 2, 632X'' is computed by cycling X' back to the frontal view domain via the same MLP structure, i.e., $X'' = \mathcal{F}'_{MLP}(X')$. To guarantee the domain consistency between X' and X'', we incorporate a cycle loss, i.e. $\mathcal{L}_{cycle}$, as expressed below.

$$\mathcal{L}_{cycle} = \|X - X''\|_1. \tag{1}$$

The benefits of the cycle structure are two-fold. First, similar to the cycle consistency based approaches [7, 54], the cycle loss can innately improve the representativeness of features, since cycling back the top-view features to the frontal view domain will strengthen the connection between both views. Second, when the discrepancy of X and X'' cannot be further narrowed down, X'' actually retains the most relevant information for view projection, since X'' is reciprocally projected from X'. Hence, X and X' refer to the features before and after view projection. X'' contains the most relevant features of the frontal view for view projection. In Fig. 3, we show two examples by visualizing the features of the front view and top view. Specifically, the way we visualize them is to select the typical channels of the feature maps (i.e., the 7th and 92nd for two examples of Fig. 3) and align them with the input images. As observed, X and X'' are similar, but quite different from X', due to the domain difference. We can also observe that, via cycling, X'' concentrates more on the road and the vehicles. X, X' and X'' will be fed into the cross-view transformer.

Cross-View Transformer (CVT). The main purpose of CVT is to correlate the features before view projection (i.e. X) and the features after view projection (i.e. X') in order to strengthen the latter ones. Since X'' contains the substantial information of the frontal view for view projection, it can be involved to further enhance the features as well. As illustrated in Fig. 4, CVT can roughly be divided into two schemes: the cross-view correlation scheme that explicitly correlates the features of views to achieve an attention map W to strengthen X' as well as the feature selection scheme that extracts the most relevant information from X''.

Particularly, X, X', and X'' serve as the key K ($K \equiv X$), the query Q ($Q \equiv X'$), and the value V ($V \equiv X''$) of CVT. In our model, the dimensions of X, X', and X'' are set as the same. X' and X are both flattened into patches, and each patch is denoted as $\mathbf{x}'_i \in X'(i \in [1, \ldots, hw])$ and $\mathbf{x}_j \in X(j \in [1, \ldots, hw])$, where hw refers to the width of X times its height. Thus, the relevance matrix R between any pairwise patches of X and X' can be estimated, i.e., for each patch $\mathbf{x}'_i$ in X' and $\mathbf{x}_j$ in X, their relevance $r_{ij}(\forall r_{ij} \in R)$ is measured by the normalized inner-product:

$$r_{ij} = \langle \frac{\mathbf{x}'_i}{||\mathbf{x}'_i||}, \frac{\mathbf{x}_j}{||\mathbf{x}_j||} \rangle. \tag{2}$$

With the relevance matrix R, we create two vectors W ($W = \{w_i\}, \forall i \in [1, \ldots, hw]$) and H ($H = \{h_i\}, \forall i \in [1, \ldots, hw]$) based on the maximum value and the corresponding index for each row of R, respectively:

$$w_i = \max_j r_{ij}, \forall r_{ij} \in R, \tag{3}$$

$$h_i = \arg\max_j r_{ij}, \forall r_{ij} \in R. \tag{4}$$

Each element of W implies the degree of correlation between each patch of X' and all the patches of X, which can serve as an attention map. Each element of H indicates the index of the most relevant patch in X with regard to each patch of X'.

Recall that both X and X'' are the features of the frontal view, except that X contains its complete information while X'' retains the relevant information for view projection. Assuming that the correlation between X and X' is similar to the correlation between X'' and X', it is reasonable to utilize the relevance of X and X' (i.e. R) to extract the most important information from X''. To this end, we introduce a feature selection scheme $\mathcal{F}_{fs}$. With H and X'', $\mathcal{F}_{fs}$ can produce new feature maps T ($T = \{\mathbf{t}_i\}, \forall i \in [1, \ldots, hw]$) by retrieving the most relevant features from X'':

$$\mathbf{t}_i = \mathcal{F}_{fs}(X'', h_i), \forall h_i \in H, \tag{5}$$

where $\mathcal{F}_{fs}$ retrieves the feature vector $\mathbf{t}_i$ from at the h_i-th position of X''.

Hence, T stores the most relevant information of X'' for each patch of X'. It can be reshaped as the same dimension as X' and concatenated with X'. Then, the concatenated features will be weighted by the attention map W and finally aggregated with X' via a residual structure. To sum

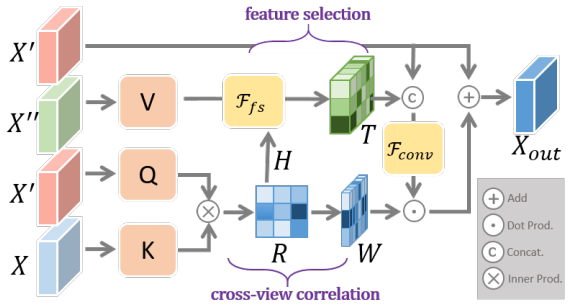


Figure 4. Illustration of cross-view transformer. It contains the cross-view correlation scheme that correlates X and X' to gain the attention map W and the feature selection scheme that extracts the most relevant information from X'' to be T.

up, the process can be formally expressed as below:

$$X_{out} = X' + \mathcal{F}_{conv}(\text{Concat}(X', T)) \odot W, \tag{6}$$

where $\odot$ denotes the element-wise multiplication and $\mathcal{F}_{conv}$ refers to a convolutional layer with 3×3 kernel size. X_{out} is the final output of CVT and will then be passed to the decoder network to produce the segmentation mask of the top view.

3.3. Context-aware Discriminator

In the discriminator of GAN-based framework, to further refine the synthetic masks of vehicles, the spatial relationship between the vehicles and their context (i.e. road) can be exploited. To accomplish this, we propose a context-aware discriminator that not only attempts to distinguish the output vehicle masks and the ground-truth ones, but also explicitly utilizes the correlation between the vehicles and the roads to strengthen the discrimination.

Particularly, with the estimated masks of vehicles $\hat{M}_v$ and the ground-truth mask of the road M_r in the same scene, we deploy a shared CNN $\mathcal{F}_D$ to separately extract the features of $\hat{M}_v$ and the concatenation of $\hat{M}_v$ and M_r, and then calculate the inner-product of their features to evaluate their correlation, i.e.,

$$\hat{C}_{v,r} = \langle \mathcal{F}_D(\hat{M}_v), \mathcal{F}_D(\{\hat{M}_v, M_r\}) \rangle. \tag{7}$$

Likewise, the ground-truth mask of vehicles M_v and the concatenation of M_v and M_r are fed through the same network with shared parameters, and then the correlation of ground-truth vehicles and road, $C_{v,r}$, is evaluated in the same way.

To this end, $\hat{M}_v$ and M_v are fed into a classifier $\mathcal{F}_D$ for a foreground object discrimination, while the correlations $\hat{C}_{v,r}$ and $C_{v,r}$ are sent into the other classifier $\mathcal{F}'_D$ for discrimination. In practice, for both classifiers, we adopt multiple convolutional layers and insert spectral normalization after each layer along with hinge losses for stabilizing training. Thus, the losses of the discriminator are:

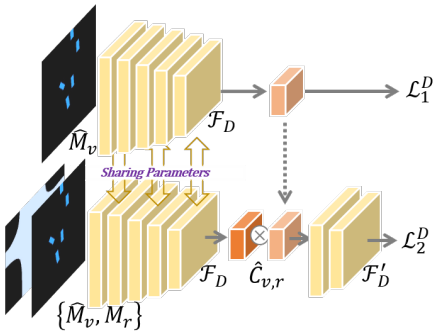


Figure 5. Illustration of the context-aware discriminator. With the estimated vehicle mask $\hat{M}_v$ and its corresponding road layout M_r, our discriminator simultaneously measures the vehicle mask and its correlation with road using the hybrid adversarial loss.

$$\mathcal{L}_1^D = \mathbb{E}[\max(0, 1 + \mathcal{F}_D(\hat{M}_v))] + \mathbb{E}[\max(0, 1 - \mathcal{F}_D(M_v))], \tag{8}$$

$$\mathcal{L}_2^D = \mathbb{E}[\max(0, 1 + \mathcal{F}'_D(\hat{C}_{v,r}))] + \mathbb{E}[\max(0, 1 - \mathcal{F}'_D(C_{v,r}))]. \tag{9}$$

Hence, our context-aware discriminator allows us to distinguish the estimated and ground-truth vehicles, meanwhile discriminating the respective correlations between the vehicles and the road, which emphasizes the spatial relationship between the vehicles and the road.

3.4. Loss Function

Overall, the loss function of our framework is defined as:

$$\mathcal{L} = \mathcal{L}_{BCE} + \lambda \mathcal{L}_{cycle} + \beta(\mathcal{L}_1^D + \mathcal{L}_2^D), \tag{10}$$

where $\mathcal{L}_{BCE}$ is a binary cross-entropy loss which serves as the main objective of the generator network to narrow the gap between the synthetic semantic mask and the ground-truth mask. λ and β are the balance weights of the cycle loss and the adversarial losses, respectively. In practice, λ and β are set as 0.001 and 1.

4. Experimental Results

To evaluate our proposed model, we conduct several experiments over a variety of challenging scenarios and against state-of-the-art methods on public benchmarks. We also perform extensive ablation experiments to delve into our network structure.

4.1. Implementation Details

We implement our framework using Pytorch on a workstation with a single NVIDIA 1080Ti GPU card. In particular, we adopt ResNet-18 [14] without bottleneck layers as our backbone. Each input of CVT utilizes one convolutional layer with kernel size 1×1. All the input images are normalized to 1024×1024 and the output size is 256×256. The

network parameters are randomly initialized and we adopt the Adam optimizer [20] and use a mini-batch size of 6. The initial learning rate is set to 1×10^{-4} and it is decayed by 0.1 after 25 epochs. In practice, our model is able to run in real time (35 FPS) on our single-GPU platform.

4.2. Datasets and Comparison Methods

Datasets. We evaluate our approach on two datasets, i.e., KITTI [12] and Argoverse [4]. Since KITTI has not provided sufficient annotation for road layout or vehicles that can be used in our task, we generally follow the practice of [27] in which the results are categorized in following datasets. For comparison with state-of-the-art 3D vehicle detection approaches, we evaluate performance on the KITTI 3D object detection (*KITTI 3D Object*) split of Chen et al. [5], i.e., 3712 training images and 3769 validation images. *KITTI Odometry* dataset is used to evaluate the road layout, whose annotation comes from the Semantic KITTI dataset [2]. Besides the previous two datasets, we evaluate the performance on *KITTI Raw* split used in [38], i.e., 10156 training images and 5074 validation images. Since its ground-truths are produced by registering the depth and semantic segmentation of Lidar scans that are not sufficiently dense, we apply image dilation and erosion to produce better ground-truth annotations. Furthermore, we also compare methods on *Argoverse* that provide a high-resolution semantic occupancy grid and vehicle detection in top view for evaluating the spatial layout of road and vehicles, with 6723 training images and 2418 validation images. For performance assessment, we adopt the mean of Intersection-over-Union (mIOU) and Average Precision (mAP) as the evaluation metrics.

Comparison Methods. For evaluation of our task, we compare our model with some of the state-of-the-art methods, including MonoOccupancy [25], MonoLayout [27], Pan et al. [32], Mono3D [5], and OFT [37]. Amongst these methods, Mono3D [5] and OFT [37] are specifically used to detect vehicles in top view. As for the quantitative results, we follow the ones reported in [27]. For MonoLayout [27], we apply their latest online reported results for comparison, which are generally better than the ones reported on their original paper. Pan et al. [32] originally adopt multiple views from different cameras to generate the top-view representation. We adapt their model for single-view input and then retrain it using the same training protocol for our task. Likewise, MonoOccupancy [25] is retrained for the benchmarks of road layout estimation, and we obtain comparable or better results than the ones reported in [27].

4.3. Performance Evaluation

Road layout estimation. To evaluate the performance of our model on the task of road layout estimation, we compare our model against MonoOccupancy [25], MonoLayout [27], and Pan et al. [31] on the datasets of *KITTI Raw* and *KITTI Odometry*. Note that, since we post-process the ground-truth annotations of *KITTI Raw*, we retrain all the comparison methods under the same training protocol. The comparison results are demonstrated in Table 1. Additionally, we also compare them on *Argoverse Road*, as shown in Table 2. As observed, in these three benchmarks, our model shows advantages over the competitors in both mIOU and mAP. Examples are shown in Fig. 6. Note that the ground-truths may contain noise, since they are converted from the measurements of Lidar. Even if so, our approach can still produce satisfactory results.

KITTI	**Raw**		**Odometry**	
Methods	mIOU (%)	mAP (%)	mIOU (%)	mAP (%)
MonoOccupancy [25]	58.41	66.01	65.74	67.84
B. Pan et al. [31]	59.58	79.07	66.81	81.79
MonoLayout [27]	66.02	75.73	76.15	85.25
Ours	**68.34**	**80.78**	**77.49**	**86.69**

Table 1. Quantitative results on *KITTI Raw* and *KITTI Odometry*.

Argoverse	**Road**		**Vehicle**	
Methods	mIOU (%)	mAP (%)	mIOU (%)	mAP (%)
MonoOccupancy [25]	72.84	78.11	24.16	36.83
B. Pan et al. [31]	71.07	86.83	16.58	39.73
MonoLayout [27]	73.25	84.56	32.58	51.06
Ours	**76.51**	**87.21**	**48.48**	**64.04**

Table 2. Results on *Argoverse Road* and *Argoverse Vehicle*.

Methods	mIOU (%)	mAP (%)
MonoOccupancy [25]	20.45	22.59
Mono3D [5]	17.11	26.62
OFT [37]	25.24	34.69
B. Pan et al. [31]	16.80	35.54
MonoLayout [27]	30.18	45.91
Ours	**38.79**	**50.26**

Table 3. Results on *KITTI 3D Object*.

Vehicle occupancy estimation. Compared with road layout estimation, estimating vehicle occupancies is a more challenging task, since the scales of vehicles vary and there exist mutual occlusions in the scenes. For evaluation, we perform comparison experiments on the benchmarks of *KITTI 3D Object* and *Argoverse Vehicle* against MonoOccupancy [25], Mono3D [5], OFT [37], MonoLayout [27], and Pan et al. [31]. The results are shown in Table 2 and 3. In Table 3, our model demonstrates superior performance against the comparison methods. Since *KITTI 3D Object* contains several challenging scenarios, most comparison methods can hardly obtain 30% mIOU, while our model gains 38.79%, which shows at least 28.5% improvement over prior methods. For the evaluation on *Argoverse Vehicle*, our model outperforms others by a large margin, i.e., at least 48.8% and 25.4% boost over the comparison methods in mIOU and mAP. Note that, the dataset *Argoverse* provides the vehicle occupancies and the corresponding road layout. Thus, it is not only our cross-view trans-

Figure 6. Comparison results of road layout estimation on *KITTI Odometry*.

former but also our context-aware discriminator that plays an important role. On the first three rows of Fig. 7, we show the examples on vehicle occupancy estimation on *KITTI 3D Object*. For the challenging cases with multiple vehicles parking on the side of roads, our model can still perform well. On the last four rows of Fig. 7, we show the examples of the joint estimation for road and vehicles on *Argoverse* and we highlight the advantages of our results.

4.4. Ablation Study

To delve into our network structure, we conduct several ablation experiments with regard to the cross-view transformation module and the context-aware discriminator.

Cross-view transformation module. Recall that our cross-view transformation module consists of CVP and CVT. Specifically, CVP can be divided into the MLP and the cycle structure. CVT can be decomposed into a cross-view correlation scheme and a feature selection scheme. In the following, we will investigate the necessity of these modules, based on the dataset *KITTI 3D Object* in Table 4.

First of all, the baseline is the vanilla encoder-decoder network using the same encoder and decoder as our model. Then, we insert the MLP structure to the baseline. As shown in Table 4, it obviously improves the effectiveness of view projection. Next, we add a cross-view correlation scheme into the network, which measures the correlation of X and X' and applies it as the attention map to enhance X'. As observed, with the involvement of the cross-view correlation scheme, the performance is significantly boosted. After that, we introduce the cycle structure as well as the cycle loss into the network, in which X'' will be fed into the cross-view correlation scheme. Finally, we insert the feature selection scheme, which further strengthens the model's performance.

Structure	mIOU (%)	mAP (%)
Baseline	22.31	34.58
+ MLP	27.42	37.44
+ Cross-view correlation	35.03	46.33
+ Cycle structure	35.54	47.29
+ Feature selection	38.79	50.26

Table 4. Effectiveness of the cross-view transformation module.

K	V	mIOU (%)	mAP (%)
X'	X'	25.43	40.82
X''	X''	27.71	41.39
X	X	35.13	44.25
X''	X	37.58	49.50
X	X''	38.79	50.26

Table 5. Different input combinations of the cross-view transformer evaluated in *KITTI 3D Object*.

Cross-view transformer. We validate different input combinations of K, Q, V for CVT. We demonstrate the results in Table 5. For all test cases, the query (i.e. Q) is assigned with the feature after view projection X'. As the most trivial case, we use X' as K and V of CVT as well, which self-correlates all the non-local patches of X'. Since X' may lose some information via view projection, CVT does not perform well. Considering both K and V are assigned with X or X'', it involves the features before view projection, but X contains richer information than X'', which leads to better performance. Moreover, with X and X'' corresponding to K and V, the substantial information for view projection is implicitly introduced by X'' to strengthen the model. More specifically, using X as the key is better to generate precise relevance embedding, while applying X'' as the value encourages the involvement of most relevant features, which thus leads to the optimal results.

Front-view MonoLayout Ours Ground-truth

Figure 7. Vehicle occupancy estimation results on *KITTI 3D Object* and the joint estimation on *Argoverse*.

Context-aware discriminator. We show the effectiveness of context-aware discriminator in Table 6 and we apply the *Argoverse Vehicle* dataset for analysis, since it is the only dataset provides the vehicles and their corresponding road layout. We compare our model against our generator without applying any discriminator (i.e. the baseline in Table 6) and the generator paired with a standard discriminator (i.e. PatchGAN [18]). As observed in the table, our proposed discriminator achieves the best results in both terms of mIOU and mAP.

Structure	mIOU (%)	mAP (%)
Baseline	48.10	61.94
PatchGAN	47.07	62.33
Context-aware Discriminator	48.48	64.04

Table 6. Effectiveness of the context-aware discriminator evaluated in *Argoverse Vehicle* dataset.

Figure 8. We montage the estimated road layout from the image sequences of *Argoverse* to produce a panorama HD map (on the right side of the above figure) which contains road layout as well as vehicle occupancies.

4.5. Panorama HD Map Generation

We showcase the application of our model on the dataset *Argoverse* for generating panorama HD map via stitching the road layout estimation given the consecutive frontal view images. The generated HD map is shown in Fig. 8, which shows the potential of our approach of being applied for generating panorama HD map.

5. Conclusion

In this paper, we present a novel framework to estimate road layout and vehicle occupancy in top views given a front-view monocular image. In particular, we propose a cross-view transformation module that is composed of cycled view projection structure and cross-view transformer, in which the features of the views before and after projection are explicitly correlated and the most relevant features for view projection are fully exploited in order to enhance the transformed features. Besides, we propose a context-aware discriminator that takes into account the spatial relationship of vehicles and roads. We demonstrate that our proposed model can achieve the state-of-the-art performance and run at 35 FPS on a single GPU, which is efficient and applicable for real-time paranoma HD map reconstruction.

Acknowledgement. Our work is supported by the National Natural Science Foundation of China under grants (No. 62072110, 61972162, 61873067) and the Natural Science Foundation of Fujian Province under Grant 2018J07005 and CCF-Tencent Open Research fund.

References

[1] Syed Ammar Abbas and Andrew Zisserman. A geometric approach to obtain a bird's eye view from an image. In *ICCV Workshops*, pages 4095–4104, 2019. 3

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*, pages 9297–9307, 2019. 6

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3

[4] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *CVPR*, pages 8748–8757, 2019. 6

[5] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *CVPR*, 2016. 1, 2, 6

[6] Mingyu Ding, Yuqi Huo, Hongwei Yi, Zhe Wang, Jianping Shi, Zhiwu Lu, and Ping Luo. Learning depth-guided convolutions for monocular 3d object detection. In *CVPRW*, pages 1000–1001, 2020. 1

[7] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *CVPR*, pages 1801–1810, 2019. 4

[8] Rui Fan, Hengli Wang, Peide Cai, and Ming Liu. Sne-roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection. In *ECCV*, pages 340–356. Springer, 2020. 2

[9] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, pages 3146–3154, 2019. 2

[10] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity recognition. In *CVPR*, pages 839–848, 2020. 3

[11] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3d traffic scene understanding from movable platforms. *PAMI*, 36(5):1012–1025, 2014. 1, 2

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361. IEEE, 2012. 6

[13] Saurabh Gupta, Varun Tolani, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. *IJCV*, (4), 2017. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 5

[15] Joey Hong, Benjamin Sapp, and James Philbin. Rules of the road: Predicting driving behavior with a convolutional model of semantic interactions. In *CVPR*, pages 8454–8462, 2019. 1

[16] Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. Learning lightweight lane detection cnns by self attention distillation. In *ICCV*, pages 1013–1021, 2019. 1

[17] Yingping Huang, Yangwei Li, Xing Hu, and Wenyan Ci. Lane detection based on inverse perspective transformation and kalman filter. *KSII Transactions on Internet & Information Systems*, 12(2), 2018. 3

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 8

[19] Jinkyu Kim, Suhong Moon, Anna Rohrbach, Trevor Darrell, and John Canny. Advisable learning for self-driving vehicles by internalizing observation-to-action rules. In *CVPR*, pages 9661–9670, 2020. 1

[20] Diederik P Kingma and Jimmy Lei Ba. Adam: Amethod for stochastic optimization. 6

[21] Buyu Li, Wanli Ouyang, Lu Sheng, Xingyu Zeng, and Xiaogang Wang. Gs3d: An efficient 3d object detection framework for autonomous driving. In *CVPR*, 2020. 1, 2

[22] Justin Liang, Namdar Homayounfar, Wei-Chiu Ma, Shenlong Wang, and Raquel Urtasun. Convolutional recurrent network for road boundary extraction. In *CVPR*, pages 9512–9521, 2019. 1

[23] Chien-Chuan Lin and Ming-Shi Wang. A vision based top-view transformation model for a vehicle parking assistant. *Sensors*, 12(4):4431–4446, 2012. 1, 3

[24] Buyu Liu, Bingbing Zhuang, Samuel Schulter, Pan Ji, and Manmohan Chandraker. Understanding road layout from videos as a whole. In *CVPR*, pages 4414–4423, 2020. 1

[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 & Automation Letters*, 4(2):445–452, 2019. 2, 6

[26] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *AAAI*, volume 33, pages 6120–6127, 2019. 1

[27] Kaustubh Mani, Swapnil Daga, Shubhika Garg, N. Sai Shankar, Krishna Murthy Jatavallabhula, and K. Madhava Krishna. Monolayout: Amodal scene layout from a single image. In *WACV*, 2020. 2, 6

[28] Gellert Mattyus, Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Hd maps: Fine-grained road segmentation by parsing ground and aerial images. In *CVPR*, 2016. 2

[29] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *CVPR*, 2017. 2

[30] Sajjad Mozaffari, Omar Y Al-Jarrah, Mehrdad Dianati, Paul Jennings, and Alexandros Mouzakitis. Deep learning-based vehicle behavior prediction for autonomous driving applications: A review. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 1

[31] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. Cross-view semantic segmentation for sensing surroundings. In *IEEE Robotics and Automation Letters*, 2019. 3, 6

[32] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. Cross-view semantic segmentation for sensing surroundings. *IEEE Robotics and Automation Letters*, 5(3):4867–4873, 2020. 2, 6

[33] Jonah Philion. Fastdraw: Addressing the long tail of lane detection by adapting a sequential prediction network. In *CVPR*, pages 11582–11591, 2019. 1

[34] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, pages 194–210. Springer, 2020. 2

[35] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *CVPR*, pages 3501–3510, 2018. 1, 3

[36] Thomas Roddick and Roberto Cipolla. Predicting semantic map representations from images using pyramid occupancy networks. In *CVPR*, pages 11138–11147, 2020. 1, 2

[37] Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection. In *BMVC*, 2019. 1, 2, 6

[38] Samuel Schulter, Menghua Zhai, Nathan Jacobs, and Manmohan Chandraker. Learning to look around objects for top-view representations of outdoor scenes. In *ECCV*, 2018. 1, 2, 6

[39] Sunando Sengupta, Paul Sturgess, L'ubor Ladicky, and Philip Hilaire Torr. Automatic dense visual semantic mapping from street-level imagery. In *IEEE/RSJ International Conference on Intelligent Robots & Systems*, 2012. 2

[40] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *ICCV*, pages 1991–1999, 2019. 1

[41] Siddharth Srivastava, Frederic Jurie, and Gaurav Sharma. Learning 2d to 3d lifting for object detection in 3d for autonomous vehicles. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4504–4511. IEEE, 2019. 1, 2

[42] Tao Sun, Zonglin Di, Pengyu Che, Chun Liu, and Yin Wang. Leveraging crowdsourced gps data for road extraction from aerial imagery. In *CVPR*, pages 7509–7518, 2019. 1

[43] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *CVPR*, pages 2417–2426, 2019. 3

[44] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1013–1020. IEEE, 2018. 2

[45] Din Chang Tseng, Tat Wa Chao, and Jiun Wei Chang. Image-based parking guiding using ackermann steering geometry. In *Applied Mechanics and Materials*, volume 437, pages 823–826. Trans Tech Publ, 2013. 1, 3

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[47] D. Wang, C. Devin, Q. Cai, P. Krähenbühl, and T. Darrell. Monocular plan view networks for autonomous driving. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2876–2883, 2019. 2

[48] Ziyan Wang, Buyu Liu, Samuel Schulter, and Manmohan Chandraker. A parametric top-view representation of complex road scenes. In *CVPR*, 2019. 1, 2

[49] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, pages 5791–5800, 2020. 3

[50] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, pages 3684–3692, 2018. 2

[51] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, pages 325–341, 2018. 2

[52] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *CVPR*, 2017. 2

[53] Dong Zhang, Hanwang Zhang, Jinhui Tang, Meng Wang, Xiansheng Hua, and Qianru Sun. Feature pyramid transformer. In *ECCV*, 2020. 3

[54] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 4

[55] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3

[56] Xinge Zhu, Zhichao Yin, Jianping Shi, Hongsheng Li, and Dahua Lin. Generative adversarial frontal view to bird view synthesis. In *2018 International conference on 3D Vision (3DV)*, pages 454–463. IEEE, 2018. 1, 3

[57] Qin Zou, Hanwen Jiang, Qiyu Dai, Yuanhao Yue, Long Chen, and Qian Wang. Robust lane detection from continuous driving scenes using deep neural networks. *IEEE transactions on vehicular technology*, 69(1):41–54, 2019. 1