Looking Farther in Parametric Scene Parsing with Ground and Aerial Imagery

by

Raghava Modhugu, Harish Rithish Sethuram, Manmohan Chandraker, C V Jawahar

in

International Conference on Robotics and Automation
: 1
-7

Report No: IIIT/TR/2021/-1

Centre for Visual Information Technology
International Institute of Information Technology
Hyderabad - 500 032, INDIA
May 2021

Looking Farther in Parametric Scene Parsing with Ground and Aerial Imagery

*Raghava Modhugu[1] , *Harish Rithish Sethuram[1] , Manmohan Chandraker[2] , C.V. Jawahar[1]

***Abstract*— Parametric models that represent layout in terms of scene attributes are an attractive avenue for road scene understanding in autonomous navigation. Prior works that rely only on ground imagery are limited by the narrow field of view of the camera, occlusions and perspective foreshortening. In this paper, we demonstrate the effectiveness of using aerial imagery as an additional modality to overcome the above challenges. We propose a novel architecture, Unified, that combines features from both aerial and ground imagery to infer scene attributes. We quantitatively evaluate on the KITTI dataset and show that our Unified model outperforms prior works. Since this dataset is limited to road scenes close to the vehicle, we supplement the publicly available Argoverse dataset with scene attribute annotations and evaluate on far-away scenes. We show both quantitatively and qualitatively, the importance of aerial imagery in understanding road scenes, especially in regions farther away from the ego-vehicle. All code, models, and data, including scene attribute annotations on the Argoverse dataset along with collected and processed aerial imagery, are available at https://bit.ly/2QsKNeR.**

I. Introduction

The ability to understand complex road scenes plays a crucial role in autonomous navigation and is an active area of research. Semantic understanding of the world can be obtained with non-parametric methods like semantic segmentation [1], [2], [3], [4] and depth estimation [5], [6], [7]. However, despite accurate semantics, non-parametric outputs do not correspond to typical human interpretations associated with driving, thus, might not be intuitive to use for downstream navigational reasoning or decision making tasks.

In contrast to the above approaches, Wang et al [8] and Liu et al [9] propose a rich parametric model to represent the 3D scene layout from monocular ground imagery, which facilitates high-level reasoning. However, the model faces challenges in representing prominent aspects in the scene layout outside the field of view of the camera and in estimating distance and semantics due to perspective foreshortening. Refinement, extraction and understanding of global road topology [1], [10], [11], [12], [13] using aerial imagery is advantageous due to its larger field of view, uniform resolution for distance and semantic estimation in near and far fields (due to a nearly orthographic projection). However, it cannot observe all local properties of the road topology due to occlusions in aerial imagery arising from vegetation and infrastructure. Given the complementary properties of ground and aerial data, we propose that benefits may be available through their combination. However, this is a challenging problem for scene understanding, since both local details and global information are manifested very differently in perspective and aerial imagery.

We propose to use both, aerial and ground modalities, for the task of parametric representation of road scenes to exploit their complementary properties as illustrated in Figure 1. Several prior works exploit such complementary understanding from multiple data modalities such as ground and aerial imagery [14], [15], [16], [17], [18], [19], [20], [21] , ground imagery and LIDAR data from perspective view [22], [23], ground imagery and Open Street Maps (OSM) [22], aerial imagery and OSM [22], as well as LIDAR and OSM [22], [24], [25]. We differ from all those works in using aerial imagery to obtain global context and ground imagery for strong visual cues of local measurements, in order to derive a parametric representation of scene geometry and semantics in unconstrained traffic scenes. We propose a new scene model, network architecture and multi-task training strategy to fully realize the complementary benefits of ground and aerial imagery.

The parameters used to define our scene layout are existence and distance to intersection, type of intersection, number of lanes on both sides, number of lanes on opposite side etc. We use the annotations released by [26] on the KITTI dataset for comparison of results with previous works. This dataset is however, limited to describing scenes that are only up to 30 meters from the ego-vehicle. Thus, we supplement the existing Argoverse dataset [27] with scene attributes annotations that can describe road scenes that are much farther away from the ego-vehicle. It is important to note that manual annotation for these parameters is very subjective, needs complex reasoning of the geometry of the scene and expensive to collect at large scale. We thus leverage the publicly available Argoverse HD map [27] and automatically extract attributes from it. We then show the increasing importance of aerial imagery in predicting scene attributes that rely on distant visual cues.

To summarize, in this paper, we make the following contributions:

- A novel approach to parametric road scene understanding that leverages complementary properties of ground and aerial imagery.
- Novel representations that yield advantages with respect to field of view, occlusions and estimation farther from the ego-vehicle.

* Both authors contributed equally to this research.

[1]The authors are with Center for Visual Information Technology(CVIT), IIIT Hyderabad, India. durga.nagendra@research.iiit.ac.in, harishrithish7@gmail.com, jawahar@iiit.ac.in.

[2] The author is with University of California, San Diego. mkchandraker@eng.ucsd.edu

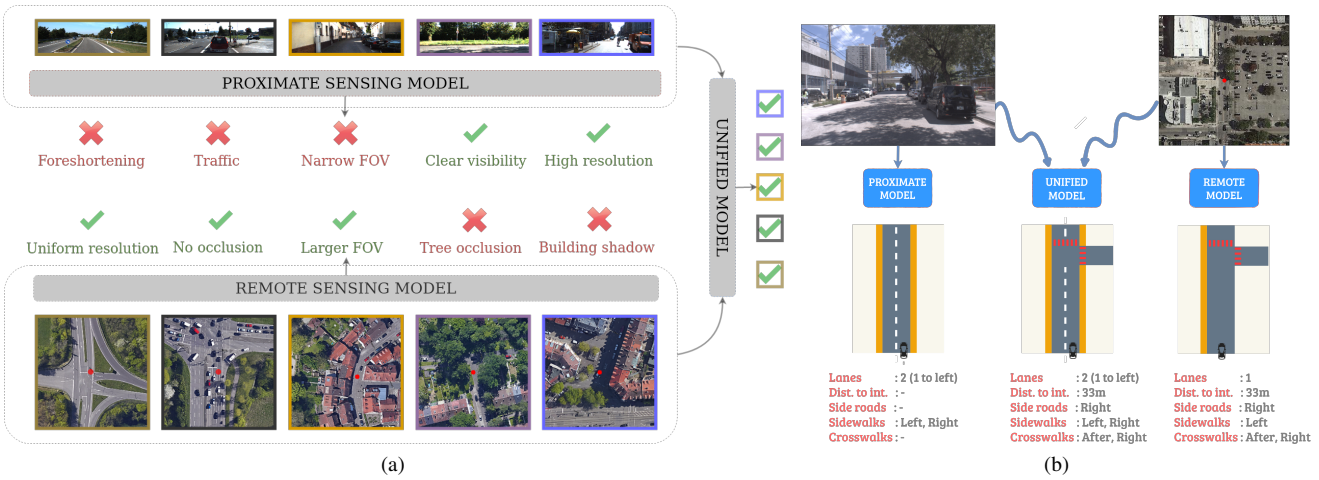


Fig. 1: (a) While ground imagery provides strong cues for local properties of the road topology, they are limited by the narrow field of view of the camera, occlusions and perspective foreshortening. In contrast, aerial imagery has the advantage of a larger field of view, presents a uniform resolution in both the near and far fields and is free from severe occlusions due to traffic. (b) In this paper, we derive a parametric representation of scene geometry and semantics in unconstrained traffic scenes. We leverage aerial imagery to lookahead and complement the local visual cues from ground imagery. Note that the bright red spot on the aerial image corresponds to the position of the ego-vehicle.

- A dataset with processed aerial imagery and scene attribute annotations to supplement the publicly available Argoverse dataset.

II. Related Work

a) Parametric scene understanding: Parametric scene understanding is the task of approximating a road scene with a set of parameters, that are often human interpretable and thus intuitive to use for downstream navigational reasoning or decision making tasks. Ess et. al. [2] propose a method to distinguish between different road layouts and also detect the presence of cars and pedestrians, while Mattyus et al. [11] estimate the width of OSM roads by utilizing aerial imagery. These methods only provide a crude understanding of the road scene. Geiger et al. [28] reason about the scene topology, geometry, and traffic activities from hand-crafted features and evaluate on a limited dataset of 113 images. Mattyus et al. [14] jointly infer location and width of roads, cycling pavements, parking areas and sidewalks using aerial and stereo ground imagery. They use handcrafted features which model only straight roads and do not work on scenes with intersections. In contrast, our work can model complex road scenes including intersections.

We find [8], [9], [29] to be the closest works to ours. Seff et al. [29] use CNN to automatically infer scene attributes from a monocular RGB image. Wang et al. [8] estimate the semantic layout of ground imagery in Birds Eye View (BEV) and then extract scene attributes. Liu et al. [9] propose to use videos to benefit from cues of camera motion and long-term context. They make use of a Feature Transform Module to fuse features from nearby frames and a COLMAP [30], [31] based scene reconstruction of the whole video sequence to provide global information. However, all these models still suffer from issues of occlusion and perspective foreshortening inherently present in ground imagery, which we overcome by providing imagery from aerial modality. Additionally, we are able to provide surrounding scene context to the model by simply enlarging the field of view in aerial imagery.

b) Aerial-ground reasoning: Hu et al. [33] address the task of geo-localizing a ground-view image on an aerial image. Li et al. [34] propose a method to adapt to ground imagery in unseen regions with the help of aerial imagery. While these methods utilize both modalities of data, they focus on obtaining similar features from both views. In contrast, we leverage the complementary properties of the modalities to obtain stronger representations. Wegner et al. [35] detect trees by jointly reasoning from aerial and ground imagery. They use per-view detectors to obtain detection proposals from each image separately. The proposals from each view are then combined to generate the final proposals. While they combine the two modalities at the output-level, we combine the aerial and ground imagery at the feature-level. This provides the model stronger cues to learn and predict from. Feng et al. [36] propose a method for urban zoning by using the semantic information from aerial imagery and the classification outputs of ground imagery. The extracted features are then passed to an MRF for inference. However, they use handcrafted features and their method is not trainable end-to-end. Workman et al. [17] propose an end-to-end trainable network for estimating geospatial functions such as population density, land cover and land use. They use kernel regression and density estimation to convert features from ground-level images into a dense feature map and combine them with the aerial imagery features. Manderson et. al [18] learn a navigation policy for off-road driving leveraging complementary inputs of ground and aerial imagery. In comparison, we learn novel representations that

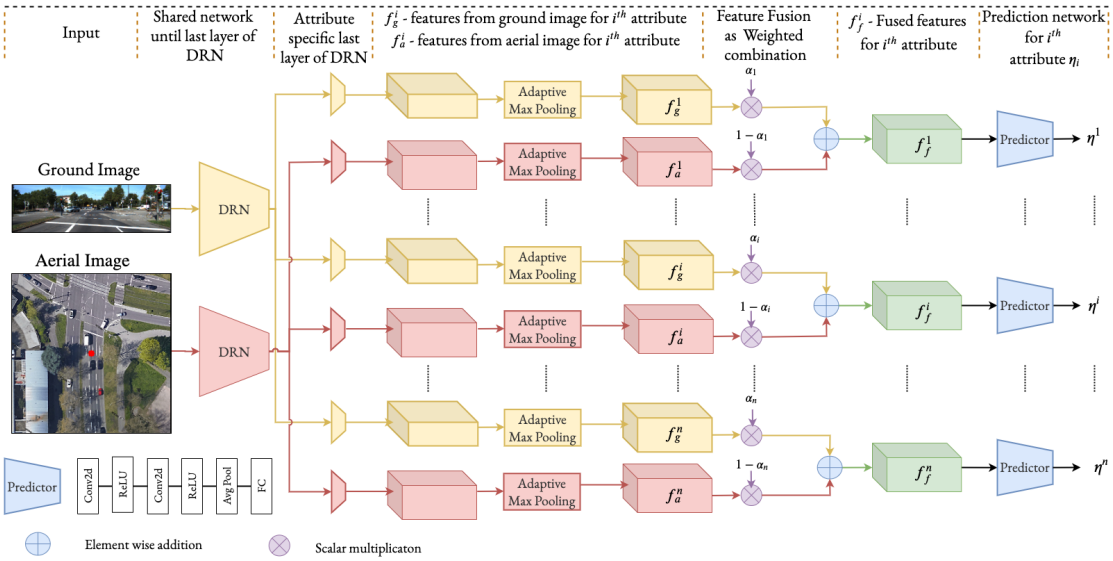


Fig. 2: We employ two Dilated Residual Networks [32] (DRN), till the penultimate layer, to extract generic features from ground and aerial imagery. The ultimate layer of the DRN is then used to extract targeted features for each attribute, individually. The attribute-specific aerial and ground features are then passed through an adaptive max pooling layer and fused using learned weights (α_i). The fused features are passed to a convolution neural network to predict binary, continuous and multi-class attributes ($\eta_1 - \eta_n$) where n is the number of attributes. Note that the bright red spot in the aerial image corresponds to the position of the ego-vehicle.

provide an accurate representation of scene geometry and semantics in unconstrained traffic scenes. Our architecture yields advantages with respect to field of view, occlusions and estimation farther from the ego-vehicle.

III. Parametric Scene Parsing with Ground and Aerial Imagery

a) Scene Model: We can model a complex road scene using a set of binary, continuous and multi-class attributes. Binary attributes indicate the presence or absence of road components such as neighboring lanes, intersections and side roads. Mutli-class and continuous attributes provide detailed understanding of these components by quantifiying them, for example, the number of neighboring lanes and the distance to intersection. In total, we have 14 binary (Θ_B), 2 multi-class (Θ_M) and 8 continuous (Θ_C) attributes for KITTI and 8 binary, 2 multi-class and 2 continuous attributes for Argoverse. For more details, refer to [26] for KITTI and Section 4 for Argoverse.

A. Architecture

As discussed earlier, each driving scene is described by a rich set of attributes. Since each attribute describes a different component of the road, naturally, the visual cues attended to in the image vary significantly. With the aerial and ground modalities providing complementary properties, it is important to fuse features efficiently to leverage this advantage. Additionally, the amount of context available from a particular modality differs for different attributes. Finally, due to the large number of attributes being learnt, there is an inherent trade off between increasing model capacity for learning discriminative features for each attribute and the resultant model size. These reasons make the architecture design for parametric scene understanding from complementary modalities challenging which we seek to address below.

a) Attribute specific feature extraction: We use DRN as our backbone architecture for feature extraction. The choice of the feature extractor is not a focus of our work and can be replaced by other networks like ResNet [37]. Seff et al. [29] predict scene attributes by training a separate network for each attribute, thereby wasting computation by not learning shared features. On the other extreme, Wang et al. [8] learn and predict scene attributes from a single shared branch, thereby restricting the model from learning more discriminative features per attribute. We strike a balance by learning a shared DRN from each modality until the last layer, from where we branch. This allows the model to learn attribute-specific features from each modality with minimal impact on model size.

b) Multimodal fusion for leveraging complementary properties: Efficient fusion of features from ground and aerial modalities is important to leveraging their complimentary properties since the importance of a modality varies for each attribute. For example, nearby lane information is more prominent in ground modality due to high resolution of ground imagery while information on side roads is clearer from aerial imagery due to uniform resolution of the modality. Thus we fuse the features through a weighted sum given by the equation $f_f^i = \alpha_i f_g^i + (1-\alpha_i) f_a^i$, where f_g^i, f_a^i, f_f^i are ground, aerial and fused features of the i^{th} attribute respectively and

α_i is a learnable parameter to fuse the ground and aerial features of the i^{th} attribute. Vielzeuf et al. [38] propose a multi-layer fusion approach and claim better results than the above mentioned fusion technique. However, while extending their technique to a feature extractor shared by multiple attributes, the influence of one modality on the other at the earlier layers leads to bias, resulting in the network performing well only on specific attributes. Thus, we extract features independently from each modality and then finally fuse features through weighted sum.

c) Multi-attribute prediction: The fused features for each attribute are passed to a prediction network separately. The prediction network consists of two convolutional layers, a global average pooling layer and a fully connected layer in sequence.

d) Loss function: We use weighted cross entropy and least squared error as our loss functions. As the losses of binary, continuous and multi-class are of different scales, we use multipliers to each of these losses, denoted by γ_B, γ_C and γ_M respectively.

$$\mathscr{L} = \frac{1}{N}\sum_{i=1}^{N} \gamma_B \,\mathrm{BCE}(\Theta_{B,i}, \eta_{B,i}) + \gamma_M \,\mathrm{CE}(\Theta_{M,i}, \eta_{M,i}) + \gamma_C \,\ell_2(\Theta_{C,i}, \eta_{C,i}) \tag{1}$$

where BCE is Binary Cross Entropy, CE is Cross Entropy and $\{\Theta, \eta\}_{.,i}$ denotes the i^{th} sample in the dataset of length N and corresponding scene attributes.

$\eta_B = \{\eta^1, \eta^2, \ldots, \eta^b\}$, $\eta_M = \{\eta^{b+1}, \eta^{b+2}, \ldots, \eta^{b+m}\}$, $\eta_C = \{\eta^{b+m+1}, \eta^{b+m+2}, \ldots, \eta^{b+m+c}\}$, $n = b+m+c$

where b, m, c, n are the number of binary, multi-class, continuous and total attributes respectively in the scene model.

IV. Experiments and Results

A. Datasets

We use two datasets, Argoverse and KITTI for ground imagery in perspective view. Wang et al. [26] provide scene attribute annotations for the KITTI dataset. Since the released annotations on the KITTI dataset is limited to road scenes up to 30m, we supplement the publicly available Argoverse dataset with scene attribute annotations for road scenes up to 60m. Since aerial imagery for KITTI and Argoverse are unavailable, we collect and process aerial imagery for both these datasets. For each of the above datasets, we acquire aerial imagery from Google Maps at zoom level 21. The Ground Sampling Distance (GSD) is 30cm for KITTI and 15cm for Argoverse. The aerial imagery is rotated such that the direction of the ego-vehicle always points towards the north of the aerial imagery. The scene attribute annotations on the Argoverse dataset, along with the processed aerial imagery for all datasets are released publicly. We further refer to these extended datasets on KITTI and Argoverse as KITTI-Air-PSU and Argo-Air-PSU, respectively.

a) KITTI-Air-PSU: We only use the left-front RGB images from the stereo camera in the KITTI dataset. Wang et al. [26] annotate each image with 14 binary, 2 multi-class and 8 continuous attributes. The attributes describe mainroad, intersections, sideroads, crosswalks and sidewalks. There are in total, 16273 training and 2108 validation images. Please refer to [26] for detailed information.

TABLE I: Description of attributes of the Argo-Air-PSU scene model. B: Binary, M: Multi-class, C: Continuous.

ID	Description
B1	Are there lanes to the left of the ego-vehicle?
B2	Are there lanes to the right of the ego-vehicle?
B3	Is the main road one-way?
B4	Is there a side road to the left at the next intersection?
B5	Is there a side road to the right at the next intersection?
B6	Is the ego-vehicle at an intersection?
B7	Does the main road continue after the next intersection?
B8	Is the main road curved?
M1	Number of lanes to the left of the ego-vehicle?
M2	Number of lanes to the right of the ego-vehicle?
C1	Distance to the next intersection
C2	Radius of curvature

b) Argo-Air-PSU: The KITTI-Air-PSU dataset is limited to describing scenes that are only up to 30 meters from the ego-vehicle. In an another work, Seff and Xiao [29] release annotations for 1 million Google Street View (GSV) panoramas, by automatically extracting attributes from the crowdsourced Open Street Maps (OSM). While this dataset is huge, there are several drawbacks associated with it. While there is a severe misalignment between GSV imagery and the roads in OSM, the annotations also are incomplete and error-prone due to the weak vetting process. In contrast, Argoverse provides HD maps from which we automatically query and extract accurate scene attribute annotations. We use the center-front images in the Argoverse tracking dataset as the ground imagery and create two different versions of the dataset. The first version covers scene attributes that are up to 30 meters in front of the camera, which we further call Argo-Air-PSU-30. The second version covers scene attributes that are up to 60 meters in front of the camera, which we further call Argo-Air-PSU-60. Since the Argoverse dataset only contains information pertaining to roads, we annotate the Argo-Air-PSU dataset on mainroads, intersections and sideroads. In total, we obtain annotations for 13122 training and 5017 validation images with 8 binary, 2 multi-class and 2 continuous attributes as described in Table I.

c) Evaluation Metrics: For binary and multi-class attributes, we report on binary accuracy (Accu-Bi.) and multi-class accuracy (Accu-Mc.) respectively. Since few attributes in the dataset are highly biased, we experiment with F1 scores. However, since a single misprediction on the minor class results in heavy penalization, we observe that accuracy better reflects the performance of the model. For continuous attributes, we report on the normalized MSE (nMSE) scores. We do report on IOU as the renderer that is required for computation is imperfect and the IOU scores are influenced by the implementation of the renderer.

TABLE II: Comparison of our Unified model with online methods on the KITTI, Argo-Air-PSU-30 and Argo-Air-PSU-60 validation sets. Across datasets and attribute types, the Unified model shows better performance than prior works [8], [29] and baselines that use only a single modality.

Method	KITTI [28]			Argo-Air-PSU-30			Argo-Air-PSU-60		
	Accu-Bi. ↑	Accu-Mc. ↑	nMSE ↓	Accu-Bi. ↑	Accu-Mc. ↑	nMSE ↓	Accu-Bi. ↑	Accu-Mc. ↑	nMSE ↓
M-RGB [29]	.811	.778	.230	-	-	-	-	-	-
M-BEV [8], [39]	.820	.777	.141	-	-	-	-	-	-
Proximate (ours)	.833	.795	.168	.885	.838	.089	.883	.842	.088
Remote (ours)	.817	.796	**.116**	.93	.792	.058	.893	.811	.062
Unified (ours)	**.848**	**.819**	.118	**.939**	**.896**	**.047**	**.904**	**.852**	**.052**

B. Implementation Details

We set the batch size to 12 and use Adam for optimization with a learning rate of 10^{-4}. We set multipliers for binary, continuous and multi-class losses to $\gamma_B = 30$, $\gamma_C = 0.01$ and $\gamma_M = 30$, so that their respective losses lie in the same scale. We train every model on 4 NVIDIA 1080 GPUs for 10 epochs. The code and the models are released publicly.

C. Baselines

To the best of our knowledge, [8], [9] and [29] are the only works that perform parametric road scene understanding. We note that recent works like [4], [39] generate semantic maps or occupancy grids in BEV, however, the outputs are non-parametric and thus, not directly comparable to us.

M-RGB [29] The features are extracted from ground imagery using a shared ResNet-101 [37]. The scene attributes are inferred directly by passing features through a fully connected (FC) network.

M-BEV [8], [39] The model constructs the BEV of ground imagery using semantic and depth labels. Features are extracted from BEV using CNN and followed by FC network for predictions.

Liu et al. [9] The model takes a video sequence from ground modality, converts them to BEV and fuses features from different frames using a Feature Transform Module. Further, an expensive COLMAP based reconstruction applied on the entire video sequence in an offline manner is provided as an additional input to the network.

Proximate (ours) We use the Unified model proposed by us but only use features from ground imagery.

Remote (ours) We use the Unified model proposed by us but only use features from aerial imagery.

D. Results

a) Comparison on KITTI dataset (online methods): In comparison to prior works, Table II shows our Unified model achieving a significant improvement on binary accuracy (.820 to .848), multi-class accuracy (.777 to .819) and on nMSE scores (.141 to .118) for continuous attributes. Comparing Proximate with M-RGB, where the only difference is the architecture being used, we can clearly observe the impact of our design choices. Despite M-BEV using additional semantic and depth information, our Remote model still performs better on multi-class and continuous attributes, while there is a minor performance drop on binary attributes. We now compare the performance of our Unified model against our Proximate and Remote models. While the Unified model shows similar performance to Remote on nMSE scores, there is a marked increase in binary and multi-class accuracy over both Remote and Proximate models. Overall, the performance of the Unified model clearly shows the advantage of using both aerial and ground modalities for parametric scene understanding.

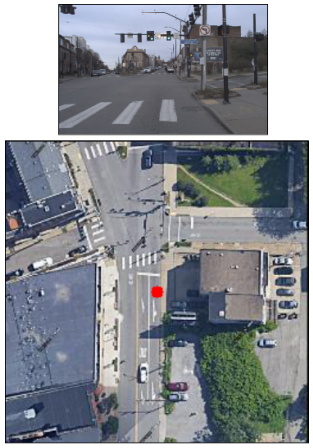
TABLE III: Comparison of our Unified model with offline methods on the KITTI validation set.

Method	KITTI [28]		
	Accu-Bi. ↑	Accu-Mc. ↑	nMSE ↓
Liu et al. [9]	.842	**.841**	.134
Unified (ours)	**.848**	.819	**.118**

b) Comparison on Argo-Air-PSU-30 and Argo-Air-PSU-60 datasets (online methods): As shown in Table II, the results indicate that the Unified model shows better performance than Remote and Proximate models on binary, multi class and continuous attributes. Importantly, they demonstrate the ability of aerial imagery in looking ahead at scenes that are farther away from the ego-vehicle. Looking at the continuous and binary attributes, the performance of Remote model is superior to Proximate model in both the Argo-Air-PSU-30 and Argo-Air-PSU-60. Most of the binary attributes and all the continuous attributes correspond to global properties of the road topology. Thus, we can infer that addition of aerial modality improves the performance in prediction of global properties, such as details of road intersection and side roads. Similarly, we can observe that the performance of the Proximate model is better than that of the Remote model on multi-class attributes i.e. lanes to the right of ego lane and lanes to the left of ego lane (local properties). The aerial imagery also aides in improving performance of the Unified model on multi class attributes in case of occlusions in ground imagery.

c) Comparison on KITTI dataset (offline methods): From Table III, we observe that though Liu et al. [9] utilize the complete video sequence for predicting scene attributes at a particular timestep, our Unified model still performs better on binary and continuous attributes, while observing imagery

(a) Curvature of the road is visible in aerial imagery.



(b) Left sideroad is not visible in ground imagery due to limited field of view.

(c) Right sidewalk is occluded by trees in aerial imagery.

(d) Right sideroad is occluded by trees in aerial imagery.

Fig. 3: The above examples demonstrate the advantage of using both aerial and ground imagery. For all examples, the Unified model gives correct predictions. The (a) and (b) are examples where the Remote model predicts correctly while the Proximate model gives incorrect predictions. The (c) and (d) are examples where Proximate model predicts correctly while the Remote model gives incorrect predictions. The reason for incorrect predictions are mentioned in the individual captions. Example (b) is taken from the Argo-Air-PSU-30 dataset while the rest are taken from KITTI-Air-PSU-30 dataset. Note that the bright red spot on the aerial image corresponds to the position of the ego-vehicle.

only from that current timestep. We however perform worse on multi-class attributes, which constitute less than 10% of the total attributes. We note that [9] uses additional cues from scene reconstruction and vehicle localization, while we extract novel representations from aerial and ground modality without additional context and fuse them efficiently.

d) Ablation Experiments: To investigate the design choices of our Unified model, we conduct several ablation studies as shown in Table IV. Firstly, we observe that having an individual prediction network for each attribute is desirable,

TABLE IV: The table shows the results of various ablation studies performed with the Unified model on the Argo-Air-PSU-30 dataset. Each row block corresponds to an experiment set. The results are to be compared within the block and with the final row, corresponding to our final Unified model. GAP: Global Average Pooling, AMP: Adaptive Max Pooling, Uni. sum: Uniform Sum, Wt. sum: Weighted sum, Pos: Position

Car Pos.	DRN Branch	Pooling	Fusion	Prediction N/W	Argo-Air-PSU-30		
					Acc-Bi ↑	Acc-Mc ↑	nMSE ↓
Middle	No	GAP	Concat.	Shared	.922	.786	.170
Middle	No	GAP	Concat.	Individual	.924	.816	.090
Middle	Yes	AMP	Concat.	Individual	.934	.805	.070
Middle	Yes	AMP	Uni. sum	Individual	.936	.816	.085
Middle	Yes	GAP	Wt. sum	Individual	.925	.863	.099
Middle	No	AMP	Wt. sum	Individual	.935	.852	.066
Bottom	Yes	AMP	Wt. sum	Individual	.923	.821	.080
Middle	Yes	AMP	Wt. sum	Individual	**.939**	**.896**	**.047**

as the model can learn more discriminative features for prediction. Secondly, we look at different techniques for multimodal feature fusion. By learning optimal weightage for the two modalities for each attribute individually, the model using weighted sum is able to best exploit the complementary properties of the two modalities. Thirdly, we look at the pooling techniques and observe that Adaptive Max Pooling performs significantly better than Global Average Pooling since it is able to retain the spatial context of features. Fourthly, we observe that even by branching only at the final layer of the DRN leads to significant improvement, validating the importance of having attribute-specific features before fusion. Finally, by placing the car at the middle of the aerial imagery, we are able to efficiently incorporate prior context behind the ego-vehicle, thereby resulting in significant performance gains.

e) Qualitative Results: In Figure 3, we illustrate a few examples where our Unified model is able to overcome the individual shortcomings of aerial and ground imagery.

V. Conclusion

In this paper, we exploit the complementary properties of aerial and ground imagery to derive a parametric representation of scene geometry and semantics in unconstrained traffic scenes. We start by creating a dataset with scene attribute annotations to supplement the publicly available Argoverse dataset. We propose a novel approach for parametric road scene understanding and show that our Unified model achieves better performance than prior works. We also extensively show, both quantitatively and qualitatively, the advantage of jointly learning from both aerial and ground modalities. In the future, we would like to extend our work beyond parametric scene understanding to other critical tasks in autonomous navigation such as landmark localization.

Acknowledgements: This work is partly supported by DST through the IMPRINT program.

REFERENCES

[1] G. Cheng, C. Wu, Q. Huang, Y. Meng, J. Shi, J. Chen, and D. Yan, "Recognizing road from satellite images by structured neural network," *Neurocomputing*, vol. 356, pp. 131–141, 2019.

[2] A. Ess, T. Müller, H. Grabner, and L. J. Van Gool, "Segmentation-based urban traffic scene understanding." in *BMVC*, vol. 1, 2009, p. 2.

[3] S. Sengupta, P. Sturgess, L. Ladický, and P. H. S. Torr, "Automatic dense visual semantic mapping from street-level imagery," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 857–862.

[4] T. Roddick and R. Cipolla, "Predicting semantic map representations from images using pyramid occupancy networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 138–11 147.

[5] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.

[6] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 239–248.

[7] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3917–3925.

[8] Z. Wang, B. Liu, S. Schulter, and M. Chandraker, "A parametric top-view representation of complex road scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 325–10 333.

[9] B. Liu, B. Zhuang, S. Schulter, P. Ji, and M. Chandraker, "Understanding road layout from videos as a whole," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4414–4423.

[10] S. M. Azimi, P. Fischer, M. Körner, and P. Reinartz, "Aerial lanenet: Lane-marking semantic segmentation in aerial imagery using wavelet-enhanced cost-sensitive symmetric fully convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 5, pp. 2920–2938, 2018.

[11] G. Mattyus, S. Wang, S. Fidler, and R. Urtasun, "Enhancing road maps by parsing aerial images around the world," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1689–1697.

[12] N. Homayounfar, W.-C. Ma, S. Kowshika Lakshmikanth, and R. Urtasun, "Hierarchical recurrent attention networks for structured online maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3417–3426.

[13] W. Li, C. He, J. Fang, and H. Fu, "Semantic segmentation based building extraction method using multi-source gis map datasets and satellite imagery." in *CVPR Workshops*, 2018, pp. 238–241.

[14] G. Máttyus, S. Wang, S. Fidler, and R. Urtasun, "Hd maps: Fine-grained road segmentation by parsing ground and aerial images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3611–3619.

[15] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, "Predicting ground-level scene layout from aerial imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 867–875.

[16] S. Workman, S. Richard, and N. Jacobs, "Understanding and mapping natural beauty," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[17] S. Workman, M. Zhai, D. J. Crandall, and N. Jacobs, "A unified model for near and remote sensing," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2688–2697.

[18] T. Manderson, S. Wapnick, D. Meger, and G. Dudek, "Learning to drive off road on smooth terrain in unstructured environments using an on-board camera and sparse aerial images," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 1263–1269.

[19] J. Delmerico, A. Giusti, E. Mueggler, L. M. Gambardella, and D. Scaramuzza, ""on-the-spot training" for terrain classification in autonomous air-ground collaborative teams," in *2016 International Symposium on Experimental Robotics*, D. Kulić, Y. Nakamura, O. Khatib, and G. Venture, Eds. Springer International Publishing, 2017.

[20] L. Wang, D. Cheng, F. Gao, F. Cai, J. Guo, M. Lin, and S. Shen, "A collaborative aerial-ground robotic system for fast exploration," in *Proceedings of the 2018 International Symposium on Experimental Robotics*, J. Xiao, T. Kröger, and O. Khatib, Eds., 2020, pp. 59–71.

[21] J. Delmerico, E. Mueggler, J. Nitsch, and D. Scaramuzza, "Active autonomous aerial exploration for ground robot path planning," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 664–671, 2017.

[22] S. Wang, M. Bai, G. Mattyus, H. Chu, W. Luo, B. Yang, J. Liang, J. Cheverie, S. Fidler, and R. Urtasun, "Torontocity: Seeing the world with a million eyes," *arXiv preprint arXiv:1612.00423*, 2016.

[23] J. Liang and R. Urtasun, "End-to-end deep structured models for drawing crosswalks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 396–412.

[24] B. Suger and W. Burgard, "Global outer-urban navigation with openstreetmap," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1417–1422.

[25] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun, "3d traffic scene understanding from movable platforms," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 5, pp. 1012–1025, 2013.

[26] Z. Wang, B. Liu, S. Schulter, and M. Chandraker, "A dataset for high-level 3d scene understanding of complex road scenes in the top-view," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, 2019.

[27] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8748–8757.

[28] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[29] A. Seff and J. Xiao, "Learning from maps: Visual common sense for autonomous driving," *arXiv preprint arXiv:1611.08583*, 2016.

[30] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[31] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*, 2016.

[32] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 472–480.

[33] S. Hu, M. Feng, R. M. Nguyen, and G. Hee Lee, "Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7258–7267.

[34] A. Li, H. Hu, P. Mirowski, and M. Farajtabar, "Cross-view policy learning for street navigation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8100–8109.

[35] J. D. Wegner, S. Branson, D. Hall, K. Schindler, and P. Perona, "Cataloging public objects using aerial and street-level images-urban trees," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 6014–6023.

[36] T. Feng, Q.-T. Truong, D. Thanh Nguyen, J. Yu Koh, L.-F. Yu, A. Binder, and S.-K. Yeung, "Urban zoning using higher-order markov random fields on multi-view imagery data," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 614–630.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[38] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "Centralnet: a multilayer approach for multimodal fusion," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[39] S. Schulter, M. Zhai, N. Jacobs, and M. Chandraker, "Learning to look around objects for top-view representations of outdoor scenes," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 787–802.