The Right (Angled) Perspective: Improving the Understanding of Road Scenes Using Boosted Inverse Perspective Mapping

Tom Bruls*, Horia Porav*, Lars Kunze, and Paul Newman

Abstract—Many tasks performed by autonomous vehicles such as road marking detection, object tracking, and path planning are simpler in bird's-eye view. Hence, Inverse Perspective Mapping (IPM) is often applied to remove the perspective effect from a vehicle's front-facing camera and to remap its images into a 2D domain, resulting in a top-down view. Unfortunately, however, this leads to unnatural blurring and stretching of objects at further distance, due to the resolution of the camera, limiting applicability. In this paper, we present an adversarial learning approach for generating a significantly improved IPM from a single camera image in real time. The generated bird's-eye-view images contain sharper features (e.g. road markings) and a more homogeneous illumination, while (dynamic) objects are automatically removed from the scene, thus revealing the underlying road layout in an improved fashion. We demonstrate our framework using real-world data from the Oxford Robot-Car Dataset and show that scene understanding tasks directly benefit from our boosted IPM approach.

I. Introduction

Autonomous vehicles need to perceive and fully understand their environment to accomplish their navigation tasks. Hence, scene understanding is a critical component within their perception pipeline, not only for navigation and planning, but also for safety purposes. While vehicles use different types of sensors to interpret scenes, cameras are one of the most popular sensing modalities in the field, due to their low cost as well as the availability of well-established image processing techniques.

In recent years, deep learning approaches based on images have been very successful and significantly improved the performance of autonomous vehicles in the context of semantic scene understanding [1], [2]. Many of these approaches take images from a front-facing camera as their input. However, images as well as their interpretations (i.e. segmented pixels) in this perspective are often transformed into a local and/or global coordinate system (or view) to be utilized effectively within tasks such as lane detection [3], [4], road marking detection [5], road topology detection [6], [7], object detection/tracking [8]–[10], as well as path planning and intersection prediction [11], [12]. This transformation is commonly referred to as Inverse Perspective Mapping (IPM) [13]. IPM takes the frontal view as input, applies a homography, and produces a top-down view of the scene by mapping the pixels to a different 2D-coordinate frame, which is also known as *bird's-eye view*.

* equal contribution

Authors are from the Oxford Robotics Institute, Dept. Engineering Science, University of Oxford, UK. {tombruls, horia, lars, pnewman}@robots.ox.ac.uk

Fig. 1. Boosted Inverse Perspective Mapping (IPM) to improve the understanding of road scenes. *Left:* Top-down view created by applying a homography-based IPM to the front-facing image (*top*), leading to unnatural blurring and stretching of objects at further distance. *Right:* Improved top-down view generated by our Incremental Spatial Transformer GAN, containing sharper features and a homogeneous illumination, while dynamic objects (i.e. the two cyclists) are automatically removed from the scene.

In practice, IPM works well in the immediate proximity of the vehicle (assuming the road surface is planar). However, the geometric properties of objects in the distance are affected unnaturally by this non-homogeneous mapping, as shown in Fig. 1. This limits the performance of applications in terms of their accuracy and the distance at which they can be applied reliably. More crucial, however, is the effect of inaccurate mappings on the semantic interpretation of scenes, where small inaccuracies can lead to significant qualitative differences. As we demonstrate in Section V-B (Table I), these qualitative differences can manifest themselves in many ways, including missing lanes and/or late detection of stop lines (or other critical road markings).

To overcome these challenges, we present an adversarial learning approach which produces a significantly improved IPM in real time from a single front-facing camera image. This is a difficult problem which is not solved by existing methods, due to the large difference in appearance between the frontal view and IPM. State-of-the-art approaches for cross-domain image translation tasks train (conditional) Generative Adversarial Networks (GANs) to transform images to a new domain [14], [15]. However, these methods are designed to perform aligned appearance transformations and struggle when views change drastically [16]. The latter work, in which a synthetic dataset with *perfect* ground-truth labels is used to learn IPM, is closest to ours.

We demonstrate in this paper that we are able to generate reliable, improved IPM for larger scenes than in [16], which are therefore able to directly aid scene understanding tasks. We achieve this in real time using real-world data collected under different conditions with a single front-facing camera. Consequently, we must deal with *imperfect* training labels (see Section IV) created from a sequence of images and ego-motion. An Incremental Spatial Transformer GAN is introduced to address the significant appearance change between the frontal view and IPM. Compared to analytic IPM approaches our learned model is (1) more realistic with sharper contours at long distance, (2) invariant to extreme illumination under different conditions, and (3) removes dynamic objects from the scene to recover the underlying road layout. We make the following contributions in this paper:

- we introduce an Incremental Spatial Transformer GAN for generating boosted IPM in real time;
- we explain how to create a dataset for training IPM methods on real-world images under different conditions; and
- we demonstrate that our boosted IPM approach improves the detection of road markings as well as the semantic interpretation of road scenes in the presence of occlusions and/or extreme illumination.

II. RELATED WORK

Improved IPM As indicated in Section I, many applications can be found in the literature that apply IPM. They rely on three assumptions: (1) the camera is in a fixed position with respect to the road, (2) the road surface is planar, and (3) the road surface is free of obstacles. Remarkably, relatively few approaches exist that aim to improve inaccurate IPM, in case one or more of these assumptions are not satisfied.

Several works have tried to adjust for inaccuracies caused by invalidity of the first two assumptions. The authors of [17], [18] used vanishing point detection, [19] estimated the slope of the road according to the lane markings, and [20] employed motion estimation obtained from SLAM. Invalidity of the third assumption is tackled in [21] by using a laser scanner to exclude obstacles from being transformed to IPM. Another approach [22]–[24] creates a look up table for all pixels, by taking into account the distance of objects on the road surface, in order to reduce artefacts at further

distance. However, these methods generally assume simple environments (i.e. highway). Contrarily, we learn a non-linear mapping more suited for urban scenes.

Very recently, [16] proposed the first learning approach for IPM using a synthetic dataset. The authors introduced BridgeGAN which employs the homography IPM to bridge the significant appearance gap between the frontal view and bird's-eye view. In contrast, we use real-world data and consequently *imperfect* labels to generate boosted IPM for larger scenes. Therefore, our learned mapping is directly beneficial for scene understanding tasks (see Section V-B).

Semantic IPM Several methods use the semantic relations between the two views for different tasks. In [25], [26] conditional random fields in the frontal view and IPM are optimized to retrieve a coarse semantic bird's-eye-view map from a sequence of camera images. A joint optimization net is trained in [27], [28] to align the semantic cues of the two views. The authors then train a GAN to synthesize a ground-level panorama from the coarse semantic segmentation. However, because aerial images differ significantly in appearance from the ground view, there is a lack of texture and detail in the synthesized images. We generate a more detailed IPM by learning a direct mapping of the pixels from the frontal view which is more useful for autonomous driving applications.

GANs for Novel View Synthesis The rise of GANs has made it possible to generate new, realistic images from a learned distribution. In order to guide the generation process towards a desired output, GANs can be conditioned on an input image [14], [29]. Until now, these methods were restricted to perform aligned appearance transformations.

In [30], the spatial transformer module was introduced to learn transformations of the input to improve classification tasks. The authors of [31], [32] used similar ideas to synthesize new views of 3D objects or scenes. More recently, these two fields were combined in [33], [34]. In the latter work, realistic compositions of objects are generated for a new viewpoint. However, these techniques are limited to toy datasets or distort real-world scenes with dynamic objects.

III. BOOSTED IPM USING AN INCREMENTAL SPATIAL TRANSFORMER GAN

A. Network Overview

As a starting point, we use a state-of-the-art architecture similar to the global enhancer of [29], without employing boundary or instance maps. Additionally, as we expect a slight change in scale from the homography-based IPM image to the stitched training labels (see Section IV), we refrain from using any pixel-wise losses and instead use multi-scale discriminator losses [29] combined with a perceptual loss [35], [36] based on VGG16 [37]. While VGG16 is trained on the ImageNet [38] dataset, thus being more suitable for frontal rather than bird's-eye-view images of road scenes, we still leverage the stability of its encoded features in this study. Retraining VGG16 on bird's-eye-view images of road scenes or swapping it out for a more suitable model, may improve the quality of the generated images, but this is beyond the scope of this study.
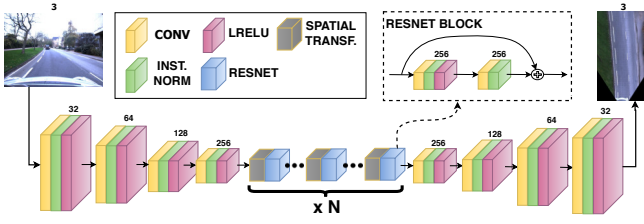
Fig. 2. The architecture of the generator of the network. The bottleneck of the model contains a series of N sequential blocks. Each block performs an incremental perspective transformation of n degrees, so that the bottleneck as a whole transforms the features from frontal to bird's-eye view. After every transformation, the features are sharpened by a ResNet block before the next transformation is applied. This process is depicted in more detail in Fig. 3.

Our model follows a largely traditional downsample-bottleneck-upsample architecture, where we reformulate the bottleneck portion of the model as a series of N_{STRes} blocks that perform incremental perspective transformations followed by feature enhancement. Each block contains a Spatial Transformer (ST) [30] followed by a ResNet layer [39]. The structure of the generator is presented in Fig. 2. For an in-depth description of the remaining architecture, the reader is directed towards the paper and supplemental material of [29].

B. Spatial ResNet Transformer

Since far-away real-world features are represented by a smaller pixel area as compared to identical close-by features, a direct consequence of applying a full perspective transformation to the input is increased unnatural blurring and stretching of the features at further distance. To counteract this effect, our model divides the full perspective transformation into a series of N_{STRes} smaller incremental perspective transformations, each followed by a refinement of the transformed feature space using a ResNet block [39]. The intuition behind this is that the slight blurring that occurs as a result of each perspective transformation is restored by the ResNet block that follows it, as conceptually visualized in Fig. 3. To maintain the ability to train our model end-to-end, we apply these incremental transforms using Spatial Transformers [30].

Intuitively, a Spatial Transformer is a mechanism, which can be integrated in a deep-learning pipeline, that warps an image using a parametrization (e.g. an affine or homography transformation matrix) conditioned on a specific input signal. Formally, each incremental spatial transformer is an end-to-end differentiable sampler, represented in our case by two major components:

- a convolutional network which receives an input I of size $H_I * W_I * C$, where H_I, W_I and C represent the height, width, and number of channels of the input respectively, and outputs a parametrization M_{loc} of a perspective transformation of size $3 * 3$, and;
- a Grid Sampler which takes I and M_{loc} as inputs, creates a mapping matrix M_{map} of size $H_O * W_O * 2$, where H_O and W_O represent the height and width of the output O. M_{map} maps homogeneous coordinates $[x, y, 1]^T$
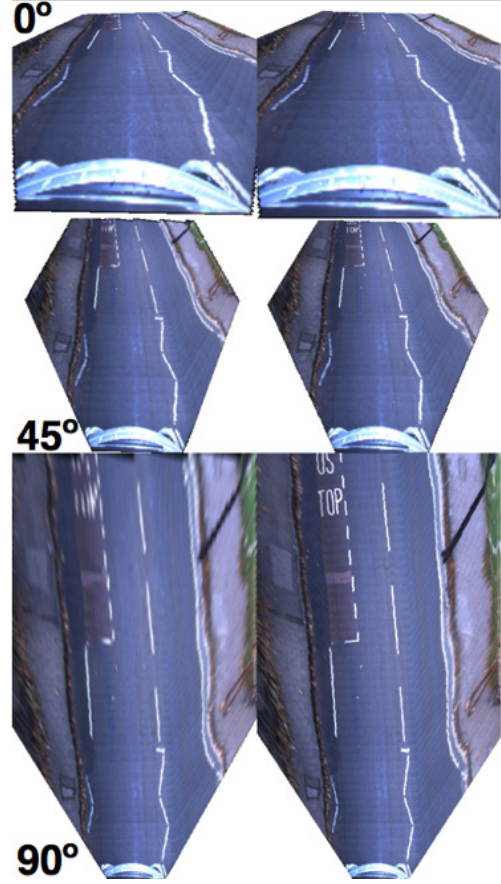

Fig. 3. Conceptual visualization of the sequential incremental transformations (i.e. $N = 3$, from $0°$ to $90°$ degrees down the rows) occurring in the bottleneck of the generator. The left column shows the features immediately after the transformation is applied, consequently they are stretched and blurred (e.g. BUS STOP letters). The right column shows how the ResNet blocks learns to sharpen these features to create the improved IPM before the next transformation is applied. Note that in reality the bottleneck has 512 feature maps instead of the 3 RGB channels depicted here for demonstration purposes.

to their new warped position given by $M_{\mathrm{loc}} * [x, y, 1]^T$. Finally, M_{map} is used to construct O in the following way: $O(x, y) = I(M_{\mathrm{map}}(x, y, 1), M_{\mathrm{map}}(x, y, 2))$.

In practice, it is non-trivial to train a spatial transformer (and even less trivial; a sequence of spatial transformers) on inputs with a large degree of self-similarity, such as road scenes. To stabilize the training procedure, for each incremental spatial transformer, we decompose $M_{\mathrm{loc}} = M_{\mathrm{locref}} * M_{\mathrm{locpert}}$, where M_{locref} is initialized with an approximate parametrization of the desired incremental homography, and M_{locpert} is the actual output of the convolutional network and represents a learned perturbation or refinement of M_{locref}.

C. Losses

Our architecture stems from [29], but does not make use of any instance maps. Due to the potential misalignment between the output of the network and the labels (see Section IV), we rely on a multi-scale discriminator loss and a perceptual loss based on VGG16. With a generator G,

k^{th} scale discriminator D_k, and $\mathcal{L}_{\text{GAN}}(G, D_k)$ being the traditional GAN loss defined over $k = 3$ scales as in [29], the final objective thus becomes:

$$\mathcal{L}_{\text{tot}} = \min_{G}((\max_{D_1, D_2, D_3} \sum_{k=1,2,3} \mathcal{L}_{\text{GAN}}(G, D_k)) +$$
$$\lambda_{\text{FM}} \sum_{k=1,2,3} \mathcal{L}_{\text{FM}}(G, D_k) + \lambda_{\text{VGG}} \mathcal{L}_{\text{VGG}}(G)), \quad (1)$$

where $\mathcal{L}_{\text{FM}}(G, D_k)$ is the multi-scale discriminator loss:

$$\mathcal{L}_{\text{FM}(G,D_k)} = \sum_{i=1}^{l_D} \frac{1}{w_i} \|D_k(I_{\text{label}})_i - D_k(G(I_{\text{input}}))_i\|_1, \quad (2)$$

and $\mathcal{L}_{\text{VGG}}(G)$ is the perceptual loss:

$$\mathcal{L}_{\text{VGG}(G)} = \sum_{i=1}^{l_P} \frac{1}{w_i} \|\text{VGG}(I_{\text{label}})_i - \text{VGG}(G(I_{\text{input}}))_i\|_1, \quad (3)$$

with l_D denoting the number of discriminator layers used in the discriminator loss, l_P denoting the number of layers from VGG16 that are utilized in the perceptual loss, and I_{input} and I_{label} being the input and label images, respectively. The weights $w_i = 2^{l-i}$ are used to scale the importance of each layer used in the loss.

D. Implementation details

We choose $N_{\text{STRes}} = 6$, $N_{\text{downsample}} = 4$, $N_{\text{upsample}} = 4$ and $l_D = l_P = 4$. Furthermore, for training, we employ the Adam solver using a base learning rate set at 0.0002, and a batch size of 1, training for 200 epochs. For the loss trade-off, we empirically set $\lambda_{\text{FM}} = 5$ and $\lambda_{\text{VGG}} = 2$. We train our network using 8416 overcast and 4894 nighttime labels. At run time, the network performs inference in real time ($\approx 20\,\text{Hz}$) using an NVIDIA TITAN X.

IV. CREATING TRAINING DATA FOR BOOSTED IPM

To evaluate our approach, we use the Oxford RobotCar Dataset [40], which features a 10-km route through urban environments under different weather and lighting conditions.

In order to create training labels which are a better representation of the real world than the standard, homography-based IPM, we use a sequence of images from the front-facing camera and corresponding visual odometry [41], and merge them into a single bird's-eye-view image.

From the sensor calibrations and the camera's intrinsic parameters, we compute the transformation which defines the one-to-one mapping between the pixels of the front-facing camera and the bird's-eye view. Then, using the relative transform obtained by visual odometry between the current image frame of the sequence and the initial frame, we stitch the respective pixels of the current frame into the IPM image at the correct pixel positions. This operation is performed iteratively, overwriting previous IPM pixels with more accurate pixels of subsequent frames, until the vehicle has reached the end of its field of view of the initial image.
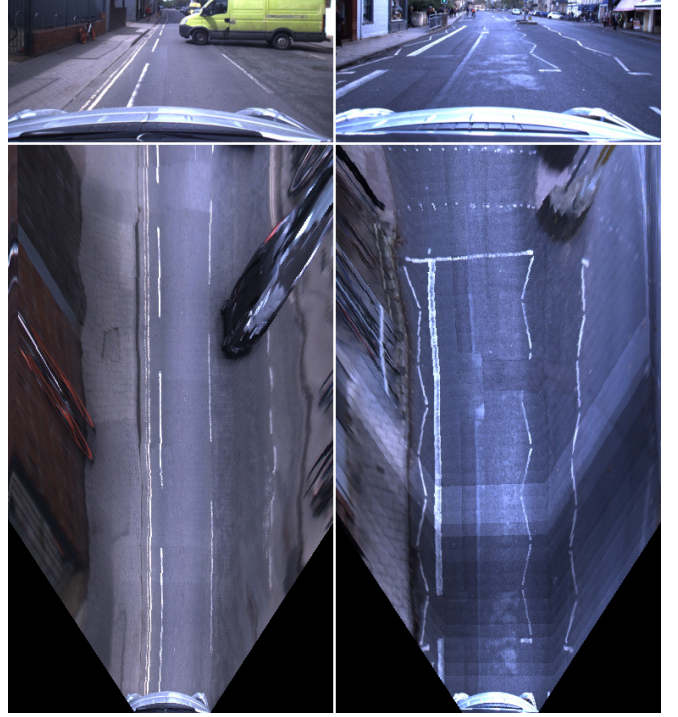

Fig. 4. Examples of created training pairs (which show the difficulties of using real-world data) by stitching IPM images generated from future front-facing camera images using the ego-motion obtained from visual odometry. The *left* example illustrates (1) movement of dynamic objects by the time the images are stitched and (2) stretching of objects because they are assumed to be on the road surface. The *right* example shows a significant change of illumination conditions. *Both* show inaccuracies at further lateral distance (e.g. wavy curb) because of sloping road surface and possibly imprecise motion estimation.

As the training labels are created from real-world data (in contrast to the synthetic data of [16]), their quality is limited by several aspects (see examples in Fig. 4):

- Minor inaccuracies in the estimation of the rotation of the vehicle and sloping road surface can lead to imprecise stitching at further lateral distance.
- Consecutive image frames may vary significantly in terms of lighting (e.g. due to overexposure), leading to illumination differences in the label which do not naturally occur in the real-world.
- Dynamic objects in the front-facing view will appear in a different position in future frames. Consequently, they will appear in unexpected places in the label.
- Objects above the road plane (e.g. vehicles, bicyclists, intersection islands, etc.) undergo a large deformation due to the view transformation. We cannot obtain accurate labels for these in real-world scenarios.

Due to the aforementioned drawbacks, no direct relation exists between the output (boosted IPM) of our network and the stitched labels. Therefore, it is impossible to incorporate a direct pixel-wise loss function, or employ super-resolution generating networks such as [42]. On the other hand, since we use a sequence of future images, regions that were previously occluded by (dynamic) objects in the initial view are potentially revealed later. This gives the network the
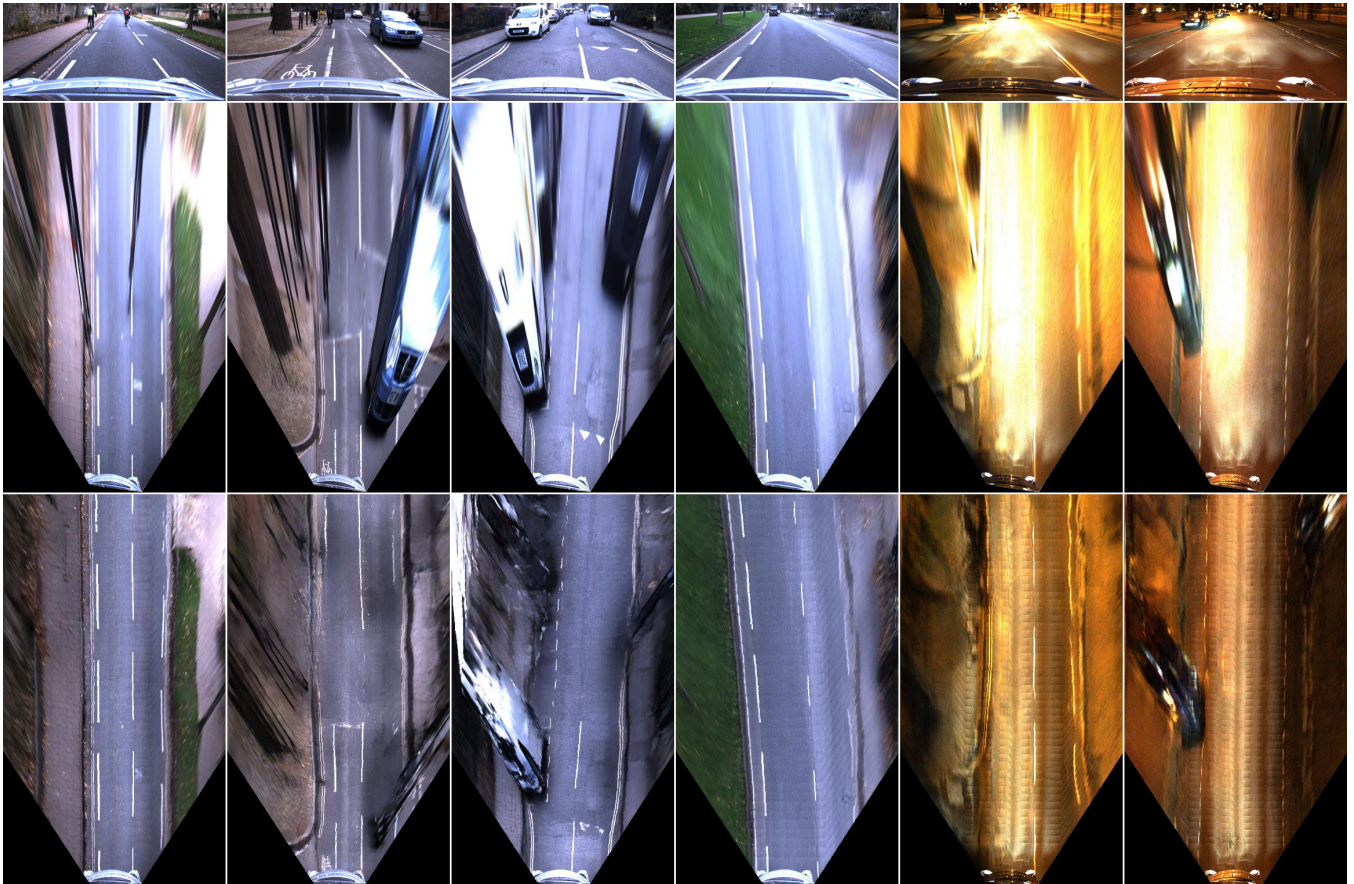
Fig. 5. Boosted IPM generated by the network (*bottom*) under different conditions compared to traditional IPM generated by applying a homography (*middle*) to the front-facing camera image (*top*). The boosted birds-eye-view images contain sharper features (e.g. road markings), more homogeneous illumination, and automatically remove (dynamic) objects from the scene. Consequently, we infer the underlying road layout, which is directly beneficial for various tasks performed by autonomous vehicles.

ability to learn the underlying road layout irrespective of occlusions or extreme illumination.

V. Experimental Results

In this section we present qualitative results generated under different conditions. Due to the nature of the problem, it is extremely hard to capture ground-truth labels in the real world (see Section IV), and thus to present quantitative results for our approach. Furthermore, the synthetic dataset used in [16] is not publicly available. However, we demonstrate that our boosted IPM has a significant qualitative effect on the semantic interpretation of real-world scenes. Lastly, we show some limitations of the presented framework.

A. Qualitative Evaluation

Fig. 5 shows qualitative results on a RobotCar test dataset. The results demonstrate that the network has learned the underlying road layout of various urban traffic scenarios. Semantic road features such as parking boxes (i.e. small separators) and stop lines are inferred correctly. Furthermore, dynamic objects, which occlude parts of the scene, are removed and replaced by the correct road/lane boundaries, making the representation more suitable for scene understanding and planning. The boosted IPM contains sharper

road markings, which improves the performance of tasks such as lane detection. Lastly, the new view offers a more homogeneous illumination of the road surface, which is beneficial for all tasks that require image processing.

Additionally, we show that our framework is not limited to datasets recorded under overcast conditions. Although artificial lighting during nighttime introduces artefacts in the output, we are still able to significantly improve the representation of the underlying layout of the scene.

B. Employing Boosted IPM for Scene Interpretation

We demonstrate the effectiveness of our improved IPM approach for the application of road marking detection [43] and scene interpretation [44] (cf. Table I). Table I shows the original front-facing camera image, the bird's-eye views (homography-based as well as our boosted IPM) and their corresponding road marking detections, and the generated graph-based scene description.

The input to the scene interpretation process is the binary image mask of the detected road markings. Within these experiments this input is either provided by the homography-based IPM or by our boosted IPM. We then cluster the road marking pixels into groups and compute a set of spatial

TABLE I
Qualitative Effects of IPM Methods on Roadmarking Detection and Scene Interpretation

Original	Road marking Detection [43]		Scene Interpretation [44]
	Homography	Boosted IPM	(generated from detected road markings)

properties and relations. Based on the spatial information and a learned probabilistic grammar, which captures the road layout of scenes, a hierarchical, graph-based scene description is generated including information about roads, lanes and road markings (which are grounded in image space). The reader is directed towards [44] for more details.

As the overall scene interpretation is based on the segmentation of road markings, the quality of the road marking detection has a major impact on the generated scene graph, as demonstrated later. Experimentally, we have verified that boosted IPM allows us to more robustly detect road markings (1) at greater distance and (2) in more detail, and (3) infer road markings occluded by dynamic objects such as cars and cyclists. These improvements are possible because boosted IPM contains sharper features with more consistent geometric properties (at further distance) and learns the underlying road layout.

We have trained a road marking detection network for each view separately (because we expect a difference in learned features) with an equivalent setup according to [43]. Labels (in the front-facing image) were generated automatically by using the techniques of [43] and mapped down into IPM to match the input images. In addition, the boosted IPM road marking labels were stitched similarly to the camera images. Although the labels are not equivalent to the ground-truth, they have proven to be sufficient for training purposes if regularization techniques are applied. The increase in performance for road marking detection in the boosted IPM has immediate consequences for the interpretation of scenes. In general, all interpretations (scene graphs) benefit from more

accurate road marking detection. Table I depicts qualitative differences in the scene graphs[1]. In the following we discuss the individual scenes.

Scene (A) The vehicle approaches a pedestrian crossing which is signaled by the upcoming zig-zag lines (visible at the top of the image). While these road markings are visible to the human eye in the homography-based IPM, the trained road marking detection network was not able to detect them because of the stretching and blurring at further distance. However, our boosted IPM produced a bird's-eye-view image with sharper contours for the zig-zag lines and correct reconstruction of the road markings occluded by the vehicle. This resulted in an improved scene graph which not only captured the right boundary of the ego lane, but also a previously undetected second lane on the right. Such qualitative differences have substantial impact on the planning and decision making of the vehicle.

Scene (B) The vehicle drives on a road with four lanes — two inner lanes for vehicles and two outer lanes for cyclists — and experiences a sudden change in illumination (from a darker foreground to a brighter background). This is clearly visible in the homography-based IPM and consequently leads to a poor detection of road markings. In contrast, our boosted approach produces a top-down view which inpaints learned semantic cues (i.e. road markings) directly over the overexposed area and also excludes the two cyclists. Hence, the resulting scene graph captures more

---

[1]In the scene graphs, the qualitative differences resulting from our boosted IPM method are indicated by filled nodes grouped in blue boxes.

detail as well as an extra lane which was missed in the segmentation resulting from the standard approach.

Scene (C) The vehicle approaches a pedestrian crossing which is indicated by both zig-zag and stop lines. Again, the distorted and blurry image resulting from the homography-based IPM leads to a poor detection of road markings. Our boosted approach has generated a more detailed view which led to better road marking detection including the successful identification of the stop lines. The resulting scene graph based on the homography-based IPM not only misses a lane, but crucially also both stop lines.

Such qualitative differences clearly demonstrate the advantage of our proposed method as they have a direct impact on planning and decision making of autonomous vehicles. While the detection and interpretation of road markings at a greater distance will enable an autonomous vehicle to adapt its behaviour earlier, the detection of road markings behind moving objects will lead to performance that is more robust and safer even when the scene is partly occluded.

C. Failure Cases

Under certain conditions, the boosted IPM does not accurately depict all details of the bird's-eye view of the scene.

As we cannot enforce a pixel-wise loss during training (Section IV), the shape of certain road markings is not accurately reflected (illustrated in Fig. 6). Improvement of the representation of these structural elements will be investigated in future work.

Furthermore, the spatial transformer blocks assume that the road surface is more or less planar (and perpendicular to the z-axis of the vehicle). When this assumption is not satisfied, the network is unable to accurately reflect the top-down scene at further distance. This might be solved by providing/learning the rotation of the road surface with respect to the vehicle.

VI. Conclusion

We have presented an adversarial learning approach for generating boosted IPM from a single front-facing camera image in real time. The generated results show sharper features and a more homogeneous illumination, while (dynamic) objects are automatically removed from the scene. Overall, we infer the underlying road layout, which is directly beneficial for tasks performed by autonomous vehicles such as road marking detection, object tracking, and path planning.

In contrast to existing approaches, we used real-world data collected under different conditions, which introduced additional issues due to varying illumination and (dynamic) objects, making it impossible to employ a pixel-wise loss during training. We have addressed the significant appearance change between the views by introducing an Incremental Spatial Transformer GAN.

We have demonstrated reliable, qualitative results in different environments and under varying lighting conditions. Furthermore, we have shown that the boosted IPM view allows for improved hierarchical scene understanding.
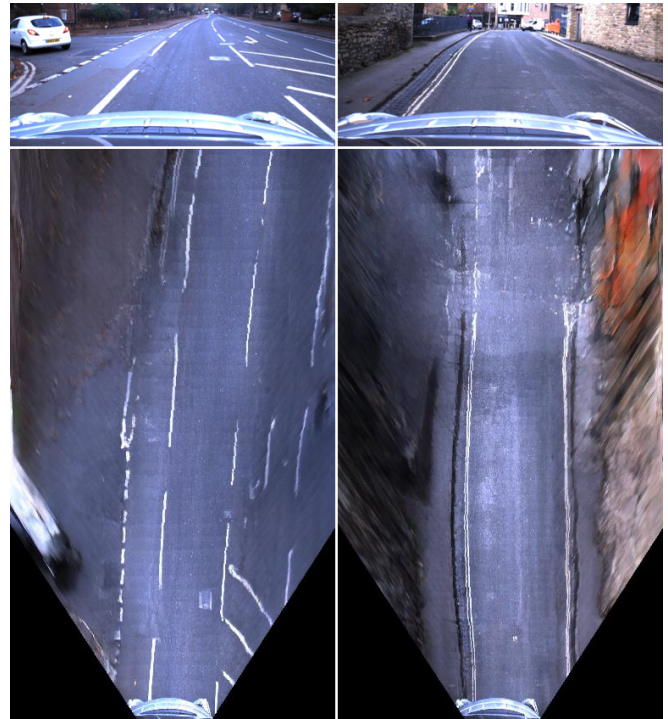

Fig. 6. Two cases in which the output of the network does not accurately depict the top-down view of the scene. In the *left* image, the road marking arrow is deformed, because we cannot employ a pixel-wise loss. In the *right* image, the road surface is not flat (sloping upwards), consequently the spatial transformer blocks attempt to map parts of the scene above the horizon, for which the features are not learned.

Consequently, our boosted IPM approach can have a significant impact on a wide range of applications in the context of autonomous driving including scene understanding, navigation, and planning.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, Dec 2017.

[2] L. Schneider, M. Cordts, T. Rehfeld, D. Pfeiffer, M. Enzweiler, U. Franke, M. Pollefeys, and S. Roth, "Semantic stixels: Depth is not enough," in *2016 IEEE Intelligent Vehicles Symposium (IV)*, June 2016, pp. 110–117.

[3] D. Neven, B. De Brabandere, S. Georgoulis, M. Proesmans, and L. Van Gool, "Towards end-to-end lane detection: An instance segmentation approach," *arXiv preprint arXiv:1802.05591*, 2018.

[4] W. Song, Y. Yang, M. Fu, Y. Li, and M. Wang, "Lane detection and classification for forward collision warning system based on stereo vision," *IEEE Sensors Journal*, vol. 18, no. 12, pp. 5151–5163, June 2018.

[5] B. Mathibela, P. Newman, and I. Posner, "Reading the road: Road marking classification and interpretation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2072–2081, Aug 2015.

[6] A. L. Ballardini, D. Cattaneo, S. Fontana, and D. G. Sorrenti, "An online probabilistic road intersection detector," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 239–246.

[7] S. Schulter, M. Zhai, N. Jacobs, and M. Chandraker, "Learning to look around objects for top-view representations of outdoor scenes," *arXiv preprint arXiv:1803.10870*, 2018.

[8] J. Dequaire, P. Ondrka, D. Rao, D. Wang, and I. Posner, "Deep tracking in the wild: End-to-end tracking using recurrent neural networks," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 492–512, 2018.

[9] N. Engel, S. Hoermann, P. Henzler, and K. Dietmayer, "Deep object tracking on dynamic occupancy grid maps using RNNs," *arXiv preprint arXiv:1805.08986*, 2018.

[10] N. Simond and M. Parent, "Obstacle detection from IPM and super-homography," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2007, pp. 4283–4288.

[11] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. S. Torr, and M. Chandraker, "DESIRE: Distant future prediction in dynamic scenes with interacting agents," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 2165–2174.

[12] A. Zyner, S. Worrall, and E. Nebot, "Naturalistic driver intention and path prediction using recurrent neural networks," *arXiv preprint arXiv:1807.09995*, 2018.

[13] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.

[14] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 5967–5976.

[15] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 2242–2251.

[16] X. Zhu, Z. Yin, J. Shi, H. Li, and D. Lin, "Generative adversarial frontal view to bird view synthesis," in *2018 International Conference on 3D Vision (3DV)*, Sep. 2018, pp. 454–463.

[17] M. Nieto, L. Salgado, F. Jaureguizar, and J. Cabrera, "Stabilization of inverse perspective mapping images based on robust vanishing point estimation," in *2007 IEEE Intelligent Vehicles Symposium*, June 2007, pp. 315–320.

[18] D. Zhang, B. Fang, W. Yang, X. Luo, and Y. Tang, "Robust inverse perspective mapping based on vanishing point," in *Proceedings 2014 IEEE International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, Oct 2014, pp. 458–463.

[19] M. Bertozzi, A. Broggi, and A. Fascioli, "An extension to the inverse perspective mapping to handle non-flat roads," in *IEEE International Conference on Intelligent Vehicles. Proceedings of the 1998 IEEE International Conference on Intelligent Vehicles*, vol. 1, 1998.

[20] J. Jeong and A. Kim, "Adaptive inverse perspective mapping for lane map generation with SLAM," in *2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, Aug 2016, pp. 38–41.

[21] M. Oliveira, V. Santos, and A. D. Sappa, "Multimodal inverse perspective mapping," *Information Fusion*, vol. 24, pp. 108 – 121, 2015.

[22] C.-C. Lin and M.-S. Wang, "A vision based top-view transformation model for a vehicle parking assistant," *Sensors*, vol. 12, no. 4, pp. 4431–4446, 2012.

[23] P. Cerri and P. Grisleri, "Free space detection on highways using time correlation between stabilized sub-pixel precision ipm images," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, April 2005, pp. 2223–2228.

[24] J. M. M. García and N. Y. Ershadi, "A new strategy of detecting traffic information based on traffic camera : modified inverse perspective mapping," *Journal of Electrical Engineering, Technology and Interface Utilities*, vol. 10, no. 2, pp. 1101–1118, March 2017.

[25] S. Sengupta, P. Sturgess, L. Ladick, 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*, Oct 2012, pp. 857–862.

[26] G. Máttyus, S. Wang, S. Fidler, and R. Urtasun, "HD maps: Fine-grained road segmentation by parsing ground and aerial images," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 3611–3619.

[27] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, "Predicting ground-level scene layout from aerial imagery," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 4132–4140.

[28] K. Regmi and A. Borji, "Cross-view image synthesis using geometry-guided conditional GANs," *arXiv preprint arXiv:1808.05469*, 2018.

[29] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 8798–8807.

[30] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[31] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee, "Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision," in *Advances in Neural Information Processing Systems*, 2016, pp. 1696–1704.

[32] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, "View synthesis by appearance flow," in *European conference on computer vision*. Springer, 2016, pp. 286–301.

[33] D. J. Rezende, S. A. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, and N. Heess, "Unsupervised learning of 3D structure from images," in *Advances in Neural Information Processing Systems*, 2016, pp. 4996–5004.

[34] S. Azadi, D. Pathak, S. Ebrahimi, and T. Darrell, "Compositional GAN: Learning conditional image composition," *arXiv preprint arXiv:1807.07560*, 2018.

[35] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[36] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 658–666.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[40] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford RobotCar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[41] W. Churchill, "Experience based navigation: Theory, practice and implementation," Ph.D. dissertation, University of Oxford, Oxford, United Kingdom, 2012.

[42] C. Ledig, L. Theis, F. Huszr, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 105–114.

[43] T. Bruls, W. Maddern, A. A. Morye, and P. Newman, "Mark yourself: Road marking segmentation via weakly-supervised annotations from multimodal data," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 1863–1870.

[44] L. Kunze, T. Bruls, T. Suleymanov, and P. Newman, "Reading between the lanes: Road layout reconstruction from partially segmented scenes," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, Nov 2018, pp. 401–408.