总结：我们提出了一个统一的端到端可训练多任务网络，共同处理在不利天气条件下消失点引导下的车道和道路标记检测和识别。为了解决这个缺点，我们建立了一条车道和道路标线基准，其中包含大约20,000张图像，包含17个车道和道路标线类，分为四种不同的场景：无雨，雨，暴雨和夜间。我们对提出的多任务网络的几个版本进行了培训和评估，并验证了每个任务的重要性。由此产生的方法VPGNet可以对车道和道路标记进行检测和分类，并通过一次正向通过来预测消失点。 我们网络的竞争优势在于它专门用于检测和识别车道和道路标记以及本地化消失点。
解决的问题： 手工制作的基于特征的方法利用边缘，颜色或纹理信息进行检测，在算法在恶劣的天气和照明条件下进行测试时，会导致性能下降。 同样，基于卷积神经网络（CNN）和手工特征相结合的方法也面临同样的挑战。 最近，已经开发了一些基于CNN的方法来以包括基于学习的算法在内的端到端方式解决该问题。 他们在基准测试和真实道路场景中表现出良好的表现，但仍然局限于晴朗的天气和简单的道路条件。VPGNet执行四项任务：网格回归，对象检测，多标签分类和消失点预测。
算法：设计了一个统一的端到端可训练多任务网络，共同处理由消失点引导的车道和道路标记检测和识别。 我们在创建的基准测试中对我们的网络进行了广泛的评估。 结果显示在不同天气条件下具有实时性能的鲁棒性。 此外，我们建议提议的消失点预测任务使网络能够检测未明确看到的车道。---> RCNN及其变体在检测和分类方面提供了突破，优于先前的方法。更快的RCNN用卷积网络替代手工建议方法，区域建议图层与分类层共享提取的要素。 表明，可以有效地计算具有滑动窗口方法的卷积网络。它也在多尺度图像的物体识别和定位方面表现出色。它的一些变体在检测任务中实现了最先进的性能。尽管这些方法在包含占据图像的重要部分的对象的大规模基准上显示尖端结果，但对于较小和较薄的物体（例如车道或道路标记），性能会下降。--->
过程：手动标注车道和道路标记的角点。 连接角点以形成多边形，从而为每个对象生成像素级掩模注释。 以类似的方式，每个像素都包含一个类标签。--->如果网络是用细线标注进行训练的，则信息往往会通过卷积和合并层消失。 此外，由于大多数神经网络需要调整大小的图像（通常小于原始大小），因此薄注释几乎不可见。因此，我们建议将像素级注释投影到网格级掩模。 如果来自原始注释的任何像素位于网格单元内，则图像被分成网格8×8并且网格单元填充有类别标签。 考虑到我们网络的输入尺寸为640×480，输出尺寸为80×60，网格尺寸被设置为与输入和输出图像之间的比例因子（1/8）成比例。具体地，网格尺寸被设置 为8×8。--->还提供了消失点注释。 我们将这个消失点定位在一个道路场景中，在这个道路场景中平行道路应该会合。 消失点由人手动注释。 根据场景的不同，难度级别（EASY，HARD，NONE）被分配给每个点。 EASY级别包括清晰的场景（例如直道）；HARD级别包括混乱的场景（例如交通阻塞）；NONE是不存在消失点的地方（例如交点）。 值得注意的是，直线和曲线都用来预测消失点。--->我们提出了一个数据层来引发网格层次的注释，使得能够同时训练车道和道路标记。我们提出了一种使用网格级掩模的替代回归方法。网格上的点被回归到最近的网格单元，并由多标签分类任务组合以表示对象。这使得能够整合具有不同特征和形状的两个独立目标，即车道和道路标记。对于后处理，车道类仅使用多标签任务的输出，路标类利用网格回归和多标签任务（。此外，我们添加了一个消失点检测任务，以在训练车道和道路标记模式期间推断全局几何背景。--->网络有四个任务模块，每个任务执行互补协作：网格框回归，对象检测，多标签分类和消失点预测。 这种结构使我们能够检测和分类车道和道路标记，并在单个正向通道中同时预测消失区域。--->我们设计了一个消失点预测（VPP）任务，用于指导类似于人类视觉的稳健车道和道路标记检测。消失点是三维空间中的平行线从图形角度收敛到二维平面的点。无论道路是弯曲的还是直线的，车道和道路标记都会聚合成一个点。在本文中，“消失点（VP）”定义为地平线上最近点，其中车道汇聚并在可见车道的最远点附近预测性地消失。该VP可用于提供场景的全局几何图形，这对于推断车道和道路标记的位置非常重要。我们将VPP模块与多任务网络相结合，将车道收敛的几何图形训练到一个点。--->作者通过使用softmax分类器矢量化网络的空间输出来预测VP的确切位置。但是，选择整个网络的输出大小中的一个点会导致不精确的定位。 为了提供更强大的本地化，我们执行多个实验来指导VP。--->附加VPP任务的目的是改善一个场景表示，这意味着一个全局上下文来预测由于遮挡或极端照明条件造成的不可见的车道。整个场景应该被考虑到，以有效反映全局信息来推断车道位置。我们使用一个象限掩模，将整个图像分成四个部分。这四个部分的交集是VP。在这方面，我们可以使用四个象限部分来推断VP，这四个部分涵盖了全球场景的结构。--->为了实现这一点，我们为VPP任务的输出定义了五个通道：一个缺少通道和四个象限通道。输出图像中的每个像素都选择属于五个通道之一。缺失通道用于表示没有VP的像素，而四象限通道代表图像上的一个象限部分。例如，如果VP存在于图像中，则应将每个像素分配给一个象限通道，而不能选择缺席通道。具体而言，第三通道将由道路场景的右上对角边缘引导，并且第四通道将从道路场景中提取左上对角边缘。另一方面，如果VP很难被识别（例如交叉路口，遮挡物），则每个像素将倾向于被分类为不通道。--->在第一阶段，我们只训练VPP任务。 除了VPP模块以外，我们将每个任务的学习率固定为零。 通过这种方式，我们可以训练内核来学习图像的全局内容。 此阶段的培训在VP检测任务达到收敛时停止。 尽管我们只训练VPP任务，但由于它们共享层的权重更新，其他检测任务的损失也减少了大约20%。 这表明车道和道路标记检测和VPP任务在特征表示层中共享一些共同的特征。

mark which consists of about 20,000 images with 17 lane and road marking classes under four different scenarios: no rain, rain, heavy rain, and night. We train and evaluate several versions of the proposed multi-task network and validate the importance of each task. The resulting approach, VPGNet, can detect and classify lanes and road markings, and predict a vanishing point with a single forward pass. Experimental results show that our approach achieves high accuracy and robustness under various conditions in real-time (20 fps). The benchmark and the VPGNet model will be publicly available [1].

1. Introduction

Autonomous driving is a large system that consists of various sensors and control modules. The first key step for robust autonomous driving is to recognize and understand the environment around a subject. However, simple recognition of obstacles and understanding of geometry around a vehicle is insufficient. There are traffic regulations dictated by traffic symbols such as lane and road markings that should be complied with. Moreover, for an algorithm to be applicable to autonomous driving, it should be robust under diverse environments and perform in real-time.

(c) (d)

Figure 1. Examples of our lane and road markings detection results in: (a) complex city scene; (b) multiple road markings; (c) night scene; (d) rainy condition. Yellow region is the vanishing area. Each class label is annotated in white.

However, research on lane and road marking detection thus far has been limited to fine weather conditions. Hand-crafted feature based methods exploit edge, color or texture information for detection, which results in a performance drop when the algorithm is tested under challenging weather and illumination conditions. Likewise, methods based on a combination of a Convolutional Neural Network (CNN) and hand-crafted features face the same challenge. Recently, a few CNN based approaches have been developed to tackle the problem in an end-to-end fashion including learning-based algorithms. They demonstrate good performance on benchmarks and in real road scenes, but are still limited to fine weather and simple road conditions.

The lack of public lane and road marking datasets is an-

[1] https://github.com/SeokjuLee/VPGNet

2380-7504/17 $31.00 © 2017 IEEE
DOI 10.1109/ICCV.2017.215

other challenge for the advancement of autonomous driving. Available datasets are often limited and insufficient for deep learning methods. For example, Caltech Lanes Dataset [1] contains 1,225 images taken from four different places. Further, Road Marking Dataset [34] contains 1,443 images manually labeled into 11 classes of road markings. Existing datasets are all taken under sunny days with a clear scene and adverse weather scenarios are not considered.

With recent advances in deep learning, the key to robust recognition in challenging scenes is a large dataset that incorporates data captured under various circumstances. Since no proper datasets available for lane and road marking recognition, we have collected and annotated lanes and road markings of challenging scenes captured in urban areas. Additionally, a higher network capability with a proper training scheme is required to generate a fine representation to cope with varied data. We propose to train a network that recognizes a global context in a manner similar to humans.

Interestingly, humans can drive along a lane even when it is hard to spot. Research works [20, 19, 28] have empirically shown that the drivers gaze direction is highly correlated with the road direction. This implies that a geometric context plays a significant role in the lane localization. Inspired by this, we aim to utilize a vanishing point prediction task to embed a geometric context recognition capability to the proposed network. Further, we hope to advance autonomous driving research with the following contributions:

- We build up a lane and road marking detection and recognition benchmark dataset taken under various weather and illumination conditions. The dataset consists of about 20,000 images with 17 manually annotated lane and road markings classes. Vanishing point annotation is provided as well.
- We design a unified end-to-end trainable multi-task network that jointly handles lane and road marking detection and recognition that is guided by the vanishing point. We provide an extensive evaluation of our network on the created benchmark. The results show robustness under different weather conditions with real-time performance. Moreover, we suggest that the proposed vanishing point prediction task enables the network to detect lanes that are not explicitly seen.

This paper is organized as follows. Section 2 covers recent algorithms developed for lane and road marking detection. A description of the benchmark is given in Section 3. Section 4 explains our network architecture and training scheme. Experimental results are reported in Section 5. Finally, Section 6 concludes our work.

2. Related Work

In this section, we introduce previous works that aim to resolve the road scene detection challenge. Our setup as well as related works is based on a monocular vision setup.

2.1. Lane and Road Marking Detection

Although lane and road marking detection appears to be a simple problem, the algorithm must be accurate in a variety of environments and have fast computation time. Lane detection methods based on hand-crafted features [9, 17, 15, 5, 29, 31, 33] detect generic shapes of markings and try to fit a line or a spline to localize lanes. This group of algorithms performs well for certain situations while showing poor performance in unfamiliar conditions. In the case of road marking detection algorithms, most of the works are based on hand-crafted features. Tao *et al*. [34] extract multiple regions of interest as Maximally Stable Extremal Regions (MSER) [25], and rely on FAST [32] and HOG[7] features to build templates for each road marking. Similarly, Greenhalgh *et al*. [13] utilizes HOG features and a SVM is trained to produce class labels. However, as in the lane detection case, these approaches show a performance drop in unfamiliar conditions.

Recently, deep learning methods have shown great success in computer vision, including lane detection. [18, 16] proposes a lane detection algorithm based on a CNN. Jun Li *et al*. [21] uses both a CNN and a Recurrent Neural Network (RNN) to detect lane boundaries. In this work, the CNN provides geometric information of lane structures, and this information is utilized by the RNN that detects the lane. Bei He *et al*. [14] proposes using a Dual-View Convolutional Neutral Network (DVCNN) framework for lane detection. In this approach, the front-view and top-view images are fed as input to the DVCNN. Similar to the lane detection algorithms, several works have examined the application of neural networks as a feature extractor and a classifier to enhance the performance of road marking detection and recognition. Bailo *et al*. [2] proposes a method that extracts multiple regions of interest as MSERs [25], merges regions that possibly belong to the same class, and finally classifies region proposals by utilizing a PCANet [6] and a neural network.

Although the aforementioned approaches provide a promising performance of lane and road marking detection using deep learning, the problem of detection under poor conditions is still not solved. In this paper, we propose a network that performs well in any situation including bad weather and low illumination conditions.

2.2. Object Detection by CNNs

With advances of deep learning, recognition tasks such as detection, classification, and segmentation have been solved under a wide set of conditions, yet there is no leading solution. RCNN and its variants [12, 11, 27] provide a breakthrough in detection and classification, outperforming previous approaches. Faster RCNN [27] replaces hand-crafted proposal methods with a convolutional network in a way that the region proposal layer shares extracted features with the classification layer. Overfeat [30] shows that a con-

volutional network with a sliding window approach can be efficiently computed. Its performance in object recognition and localization using multi-scale images is also reported. Some of its variants [23, 26] achieve state of the art performance in detection tasks. Although these approaches show cutting edge results on large-scale benchmarks [8, 10, 22], which contain objects that occupy a significant part of an image, the performance decreases for smaller and thinner objects (*e.g.* lane or road markings).

Several deep learning approaches specialize in a lane and small object recognitions. For example, Huval *et al*. [16] introduce a method for lanes and vehicles detection based on a fully convolutional architecture. They use the structure of [30] and extend the method with an integrated regression module composed of seven convolutional layers for feature sharing. The network is divided into two branches which perform binary classification and regression task. They evaluate results under a nice weather on a highway without complex road symbols, but do not perform a multi-label classification. Additionally, Zhu *et al*. [35] propose a multi-task network for traffic sign (relatively small size) detection and classification. In this work, the classification layer is added in parallel to the [16] network to perform detection and classification. As a result, this work reports better performance of detecting small objects than Fast RCNN [11].

3. Benchmark

3.1. Data Collection and Annotation

We have collected the dataset in various circumstances and categorized the images according to the time of the day and weather conditions. The dataset comprises situations during day time with different levels of precipitation: no rain, rainfall, and heavy rainfall. Night time images are not subdivided by weather condition but include general images taken in a challenging situation with low illumination. The number of images for each scenario is shown in Table 1. Since our dataset is captured under bad weather conditions, we mount a camera inside a vehicle (in the center). In this way, we can avoid damaging the camera sensor while also preventing direct water drops on the camera lens. However, since several videos are recorded in heavy rain, a part of a window wiper is captured occasionally. The camera is directed to the front view of the car. Image resolution is 1288×728. Our data are captured in a downtown area of Seoul, South Korea. The shapes and symbols of the lane and road markings follow the regulations of South Korea.

We manually annotate corner points of lane and road markings. Corner points are connected to form a polygon which results in a pixel-level mask annotation for each object. In a similar manner, each pixel contains a class label.

However, if the network is trained with a thin lane annotation, the information tends to vanish through convolution and pooling layers. Further, since most of the neu-

Table 1. Number of frames for each scenario in the dataset.

Scenario (Scn.)		Total frames	Training set	Test set
Daytime	No rain (Scn. 1)	13,925	9,184	4,741
	Rain (Scn. 2)	4,059	3,322	737
	Heavy rain (Scn. 3)	825	462	363
Night (Scn. 4)		2,288	1,815	473
Total		21,097	14,783	6,314

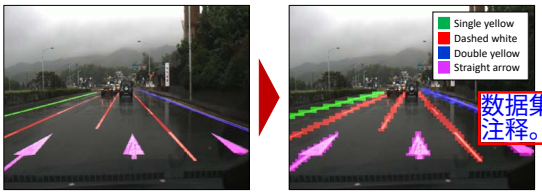


Figure 2. Pixel- and grid-level annotations of the dataset.

networks require a resized image (usually smaller than original size), the thin annotations become barely visible. Therefore, we propose projecting pixel-level annotation to the grid-level mask. The image is divided into a grid 8×8 and the grid cell is filled with a class label if any pixel from the original annotation lies within the grid cell. Considering that the input size of our network is 640×480 and the output size is 80×60, the grid size is set to be proportional to the scale factor (1/8) between the input and output images. Specifically, the grid size is set to be 8×8. Figure 2 shows an annotation example.

The vanishing point annotation is also provided. We localize the vanishing point in a road scene where parallel lanes supposedly meet. The vanishing point is manually annotated by a human. Depending on the scene, a difficulty level (EASY, HARD, NONE) is assigned to every vanishing point. EASY level includes a clear scene (*e.g.* straight road); HARD level includes a cluttered scene (*e.g.* traffic jam); NONE is where a vanishing point does not exist (*e.g.* intersection). It is important to note that both straight and curved lanes are utilized to predict the vanishing point. We describe the definition of our vanishing point in detail in Section 4.2. Furthermore, annotation examples are presented in the supplementary material.

3.2. Dataset Statistics

Our dataset consists of about 20,000 images taken during three weeks of driving in Seoul. The raw video (30 fps) is sampled at 1Hz intervals to generate image data. Images of the complex urban traffic scenes contain lane and road markings under various weather conditions during different time of the day. In total, 17 classes are annotated covering the most common markings found on the road. Although we recorded the video in various circumstances, a data imbalance between different types of lane and road markings is observed. For example, in the case of lane classes, dashed white and double yellow lines are more common than other lane types. Regarding road marking classes, straight arrows and crosswalks appear most frequently. We also define a "Other markings" class containing road markings that are

数据集中每个类的实例数

Table 2. Number of instances for each class in the dataset.

Lane		Road marking		Vanishing point	
Single white	25,354	Stop line	7,298	EASY	19,302
Dashed white	74,733	Left arrow	1,186	HARD	262
Double white	206	Right arrow	537	NONE	1,533
Single yellow	28,054	Straight arrow	6,968		
Dashed yellow	5,734	U-turn arrow	127		
Double yellow	8,998	Speed bump	1,523		
Dashed blue	1,306	Crosswalk	13,632		
Zigzag	1,417	Safety zone	6,031		
		Other markings	52,975		

present only in South Korea, or have an insufficient number of instances to be trained as a separate class. Types of classes and the number of instances are listed in Table 2.

4. Neural Network

4.1. Architecture

Our network, VPGNet, is inspired by the work of [16] and [35]. The competitive advantage of our network is that it is specialized to detect and recognize lane and road markings as well as to localize vanishing point.

We propose a data layer to induce grid-level annotation that enables training of both lane and road markings simultaneously. Originally in [16], [35], the box regression task aims to fit a single box to a particular object. This works well for objects with a blob shape (traffic signs or vehicles), but lane and road markings cannot be represented by a single box. Therefore, we propose an alternative regression that utilizes a grid-level mask. Points on the grid are regressed to the closest grid cell and combined by a multi-label classification task to represent an object. This enables us to integrate two independent targets, lane and road markings, which have different characteristics and shapes. For the post-processing, lane classes only use the output of the multi-label task, and road marking classes utilize both grid box regression and multi-label task (see Section 4.4). Additionally, we add a vanishing point detection task to infer a global geometric context during training of patterns of lane and road markings (explained in Section 4.2).

The overall architecture is described in Table 3 and Figure 3. The network has four task modules and each task performs complementary cooperation: grid box regression, object detection, multi-label classification, and prediction of the vanishing point. This structure allows us to detect and classify the lane and road markings, and predict the vanishing region simultaneously in a single forward pass.

4.2. Vanishing Point Prediction Task

Due to poor weather environments, illumination conditions, and occlusion, the visibility of lanes decreases. However, in such situations, humans intuitively can predict the locations of the lanes from global information such as nearby structures of roads or the flow of traffic [20, 19, 28]. Inspired by this, we have designed a Vanishing Point Prediction (VPP) task that guides robust lane and road marking detection similar to human vision. A vanishing point is a point where parallel lines in a three-dimensional space converge to a two-dimensional plane by a graphical perspective. In most cases of driving, lane and road markings converge to a single point regardless of whether the roads are curved or straight. In this paper, "Vanishing Point (VP)" is defined as the nearest point on the horizon where lanes converge and disappear predictively around the farthest point of the visible lane. This VP can be used to provide a global geometric context of a scene, which is important to infer the location of lanes and road markings. We integrate the VPP module with the multi-task network to train the geometric patterns of lane convergence to one point.

Borji [4] has shown that a CNN can localize the VP. The author vectorizes the spatial output of the network to predict the exact location of a VP by using a softmax classifier. However, selecting exactly one point over the whole network's output size results in imprecise localization. In order to provide more robust localization, we perform several experiments to guide the VP.

First, for the VPP task, we tried regression losses (*i.e.* L1, L2, hinge losses) that directly calculate pixel distances from a VP. Unfortunately, the results are not favorable since it is difficult to balance the losses with other tasks (object detection/multi-label classification) due to the difference in the loss scale. Therefore, we adopt a cross entropy loss to balance the gradients propagated from each of the detection tasks. By using cross entropy loss, first we apply a binary classification method that directly classifies background and foreground (*i.e.* vanishing area, see Figure 4a), as in the object detection task. The binary mask is generated in the data layer by drawing a fixed size circle centered at the VP we annotated. However, using this method on the VPP task results in extremely fast convergence of the training loss. This is caused by the imbalance of the number of background and foreground pixels. Since the vanishing area is drastically smaller than the background, the network is initialized to infer every pixel as background class. This phenomenon contradicts our original intention of training the VPP to learn the global context of a scene.

Considering the challenge imposed by the aforementioned imbalance during the binary VPP method, we have newly designed the VPP module. As stated before, the purpose of attaching the VPP task is to improve a scene representation that implies a global context to predict invisible lanes due to occlusions or extreme illumination condition. The whole scene should be taken into account to efficiently reflect global information inferring lane locations. We use a quadrant mask that divides the whole image into four sections. The intersection of these four sections is a VP. In this way, we can infer the VP using four quadrant sections which cover the structures of a global scene. To implement this, we define five channels for the output of the VPP task: one absence channel and four quadrant channels. Every pixel

Table 3. Proposed network structure.

Layer	Conv 1	Conv 2	Conv 3	Conv 4	Conv 5	Conv 6	Conv 7	Conv 8
Kernel size, stride, pad	11, 4, 0	5, 1, 2	3, 1, 1	3, 1, 1	3, 1, 1	6, 1, 3	1, 1, 0	1, 1, 0
Pooling size, stride	3, 2	3, 2			3, 2			
Addition	LRN	LRN				Dropout	Dropout, branched	Branched
Receptive field	11	51	99	131	163	355	355	355

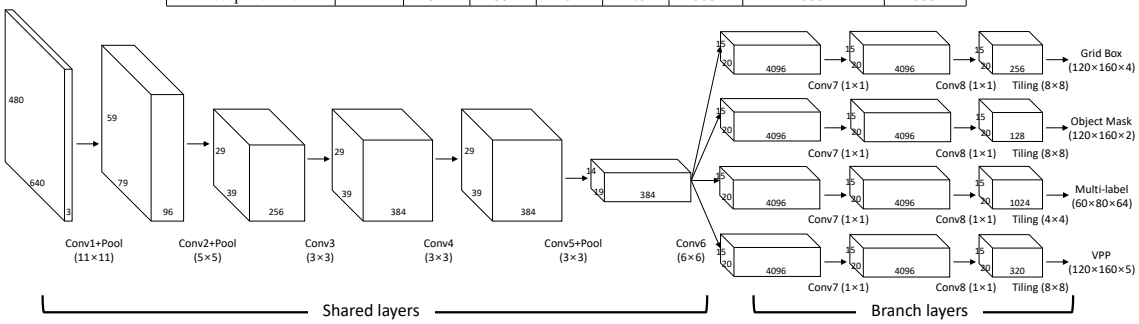


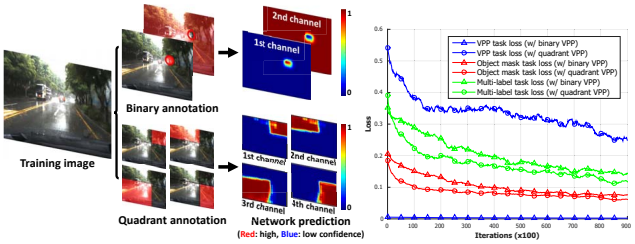
Figure 3. VPGNet performs four tasks: grid regression, object detection, multi-label classification, and vanishing point prediction.

n the output image chooses to belong to one of the five channels. The absence channel is used to represent a pixel with no VP, while the four quadrant channels stand for one of the quadrant sections on the image. For example, if the VP is present in the image, every pixel should be assigned to one of the quadrant channels, while the absence channel cannot be chosen. Specifically, the third channel would be guided by the upper right diagonal edges from the road scene, and the fourth channel would extract the upper left diagonal edges from the road scene. On the other hand, if the VP is hard to be identified (*e.g.* intersection roads, occlusions), every pixel will tend to be classified as the absence channel. In this case, the average confidence of the absence channel would be high.

Unlike the binary classification approach, our quadrant method enriches the gradient information that contains a global structure of a scene. The loss comparison in Figure 4b indirectly shows that the network is trained without overfitting compared to the binary case. Note that we only use the quadrant VPP method for the evaluation. The binary VPP method is introduced only to show readers that a naive VPP training scheme does not yield satisfactory results. The whole multi-task network allows us to detect and recognize the lane and road marking, as well as to predict the VP simultaneously in a single forward pass.

4.3. Training

Our network includes four tasks which cover different contexts. The detection task recognizes objects and covers a local context, while the VPP task covers a global context. If those tasks are trained altogether at the same training phase, the network can be highly influenced by a certain dominant task. We noticed that during the training stage the VPP task became dependent on the lane detection task. The dependency between lanes and the VP implies a strong information correlation. In this case, the VP provides redundant information to the network, leading to marginal lane detection improvement. In order to prevent this side effect, we train the network in two phases to tolerate the balance between the tasks.



(a) (b)

Figure 4. (a) Output visualization of binary and quadrant VPP methods. For the prediction of the quadrant method, only four quadrant channels are visualized except for an absence channel. (b) The loss comparison of two methods.

n the first phase, we train only the VPP task. We fix the learning rates to zero for every task except the VPP module. In this way, we can train the kernels to learn a global context of the image. The training of this phase stops upon reaching convergence of the VP detection task. Although we train only the VPP task, due to the weight update of the mutually shared layers, losses of the other detection tasks are also decreased by about 20%. This shows that lane and road marking detection and VPP tasks share some common characteristics in the feature representation layers.

n the second phase, we further train all the tasks using the initialized kernels from the first phase. Since all tasks are trained together at this point, it is important to balance their learning rates. If a certain task loss weight is small, it becomes dependent on other tasks and vice versa. Equation (1) shows the summation of four losses from each task:

$$Loss = w_1 L_{reg} + w_2 L_{om} + w_3 L_{ml} + w_4 L_{vp} \qquad (1)$$

where L_{reg} is a grid regression L1 loss, L_{om} and L_{ml} and

L_{vp} are cross entropy losses in each branch of the network. We balance the tasks by weight terms $w_1 \sim w_4$ in the following way. First, $w_1 \sim w_4$ are set to be equal to 1, and the starting losses are observed. Then, we set the reciprocal of these initial loss values to the loss weight so that the losses are uniform. In the middle of the training, if the scale difference between losses becomes large, this process is repeated to balance the loss values. The second phase stops when the validation accuracy is converged.

4.4. Post-Processing

Each lane and road marking class and VPs are required to be represented suitably for real world application. Therefore, we implement post-processing techniques to generate visually satisfying results.

Lane In the case of the lane classes, we use the following techniques: point sampling, clustering, and lane regression. First, we subsample local peaks from the region where the probability of lane channels from the multi-label task is high. The sampled points are potential candidates to become the lane segments. Further, selected points are projected to the birds-eye view by inverse perspective mapping (IPM) [3]. IPM is used to separate the sampled points near the VP. This is useful not only for the case of straight roads but also curved ones. We then cluster the points by our modified density-based clustering method. We sequentially decide the cluster by the pixel distance. After sorting the points by the vertical index, we stack the point in a bin if there is a close point among the top of the existing bins. Otherwise, we create a new bin for a new cluster. By doing this, we can reduce the time complexity of the clustering. The last step is quadratic regressions of the lines from the obtained clusters utilizing the location of the VP. If the farthest sample point of each lane cluster is close to the VP, we include it in the cluster to estimate a polynomial model. This makes the lane results stable near the VP. The class type is assigned to each line segment from the multi-labeled output of the network.

Road marking For the road marking classes, grid sampling and box clustering are applied. First, we extract grid cells from the grid regression task with high confidence for each class from the multi-label output. We then select corner points of each grid and merge them with the nearby grid cells iteratively. If no more neighboring grid cells belong to the same class, the merging is terminated. Some road markings such as crosswalks or safety zones that are difficult to define by a single box are localized by grid sampling without subsequent merging.

Vanishing point Our VPP module outputs five channels of the confidence map: four quadrant channels and one absence channel. Through these quadrants, we generate the location of a VP. The VP is where all four quadrants intersect. That is, we need to find a point where four confidences from each quadrant channel become close. Equation (2) and (3) describe the boundary intersection of each quadrant:

$$P_{avg} = \frac{1 - (\sum p_0(x,y))/(m \times n)}{4} \tag{2}$$

$$loc_{vp} = \arg\min_{(x,y)} \sum_{n=1}^{4} |P_{avg} - p_n(x,y)|^2 \tag{3}$$

where P_{avg} is the probability that a VP exists in the image, $p_n(x,y)$ is the confidence of (x,y) on n_{th} channel ($n = 0$: *absence channel*), $m \times n$ is the confidence map size, and loc_{vp} is the location of the VP.

5. Results

Our experiments consist of six parts. First, we show the experimental settings such as dataset splits and training parameters. Secondly, we provide an analysis of our network. We explore how multiple tasks jointly cooperate and affect the performance of each other. Third, our evaluation metric for each target is introduced. Lastly, we show lanes, road markings, and VPs detection and classification results.

5.1. Experimental Settings

A summary of the datasets is provided in Table 1. During the training, we double the number of images by flipping the original ones. This, in turn, doubles the training set and also prevents positional bias that comes from the lane positions. More specifically, the dataset is obtained in a right-sided driving country, and by flipping the dataset we can simulate a left-sided environment.

At the first training phase, we initialize the network only by the VPP task. After the initialization, all four tasks are trained simultaneously. For every task, we use Stochastic Gradient Descent optimization with a momentum of 0.9 and a mini-batch size of 20. Since multiple tasks must converge proportionally, we tune the learning rate of each task.

We train three models of the network divided by task: 2-Task (revised [16]), 3-Task (revised [35]), and 4-Task (VPGNet). 2-Task network includes regression and binary classification tasks. 3-Task network includes 2-Task and a multi-label classification task. 4-Task network includes 3-Task and a VPP task, which is the VPGNet. Since the lane detection in [16] is not fully reproducible, we modify the data layer to handle the grid mask and move one convolutional layer from shared layers to branch layers, as in the 3- and 4-Task networks. The 3-Task network is similar to [35], but we modify the data layer to handle the grid mask.

We test our models on NVIDIA GTX Titan X and achieve a speed of 20 Hz by using only a single forward pass. Specifically, the single forward pass takes about 30 ms and the post-processing takes about 20 ms or less.

5.2. Analysis of Multi Task Learning

In this section, we validate whether our multi-task modules contribute to improvement of the network training. We

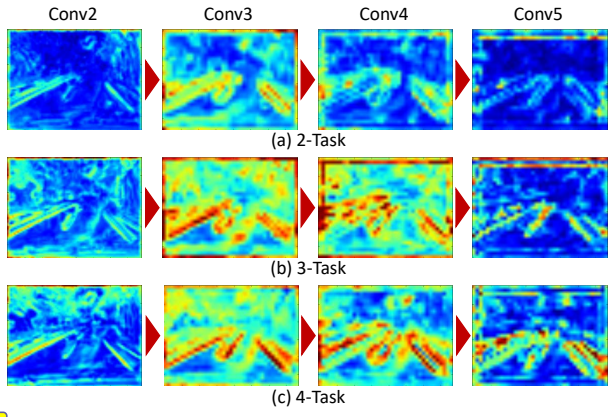


Figure 5. Activated neurons in the feature sharing network. Intensity scale in each layer activation is equalized.

observe the activated neurons in the feature sharing network. From the lower to higher layer, the abstraction level is accelerated. Figure 5 shows the activated neurons after each convolutional layer before the branch. We average over all channel values. For a fair comparison, we equalize the intensity scale in each layer activation. As the results show, if we use more tasks, more neurons respond, especially around the boundaries of roadways.

5.3. Evaluation Metrics

In this section, we show the newly proposed evaluation metrics for our benchmark evaluation. First, we introduce our evaluation metric for the lane detection. Since the ground truth of our benchmark is annotated with grid cells we compute the minimum distance from the center of each cell to the sampled lane points for every cell. If the minimum distance is within the boundary R, we mark these sampled points as true positive and the corresponding grid cell as detected. By measuring every grid cell on the lane, we can strictly evaluate the location of lane segments. Additionally, we measure F1 score for the comparison.

In the case of road markings, we use mitigated evaluation measurement. Since the only information we need while driving is the road marking in front of us rather than the exact boundary of the road markings, we measure the precision of predicted blobs. Specifically, we count all predicted cells overlapped with the ground truth grid cells. The overlapped cells are marked as true positive cells. If the number of true positive cells is greater than half of the number of all predicted cells over a clustered blob, the overlaid ground truth target is defined as detected. Additionally, we measure the recall score for comparison.

For evaluation of the VP, we measure the Euclidean distance between a ground truth point and a predicted VP. The recall score is evaluated by varying the threshold distance R from the ground truth VP. Figure 6 shows a summary of how we measure all three targets of our network.

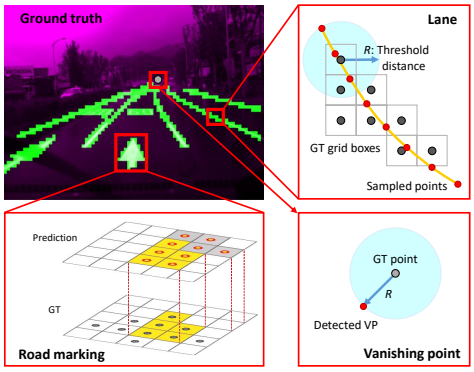


Figure 6. Graphical explanation of the evaluation metrics.

5.4. Lane Detection and Recognition

For lane classes, we measure detection, as well as simultaneous detection and classification performance. First, we compare our multi-task networks with the baseline methods in the Caltech Lanes Dataset [1] (see Figure 7). We set R to equal to the average half value of the lane thickness (20 pixels). Due to perspective effect, the double lane in front of the camera is about 70 to 80 pixels thick, and it is as small as 8 pixels (a single grid size) near the VP. Since this dataset contains relatively easy scenes during daytime, the overall performance of 2-, 3-, and 4-Task networks is very similar. Nevertheless, our network achieves the best F1 score.

Further, we provide a comparison of the proposed three versions of multi-task networks and the FCN-8s [24] segmentation method on our benchmark dataset. It is important to note that our networks utilize grid-level annotation, while FCN-8s is trained independently with both pixel- and grid-level annotations. For testing purposes, four scenarios have been selected as in Section 5.1, and the F1 score is compared in each scenario. Figure 8 shows the experimental results. Noticeably, our method shows significantly better lane detection performance in each bad weather condition scenario. Moreover, the forward pass time of the VPGNet is 30 ms, while FCN-8s [24] takes 130 ms.

Interestingly, FCN-8s shows better performance with the proposed grid-level annotation scheme compared to pixel-level annotation. This proves that the grid-level annotation is more suitable for lane detection and recognition. The reason is that grid-level annotation generates stronger gradients from the edge information around the thinly annotated area (*i.e.* lane or road markings), which, in turn, results in en-

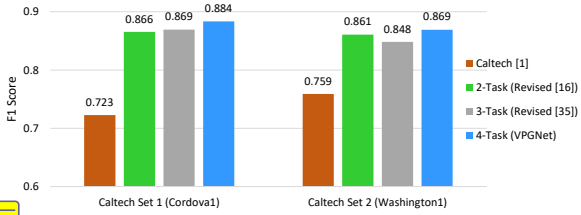


Figure 7. Lane detection score on Caltech lanes dataset.

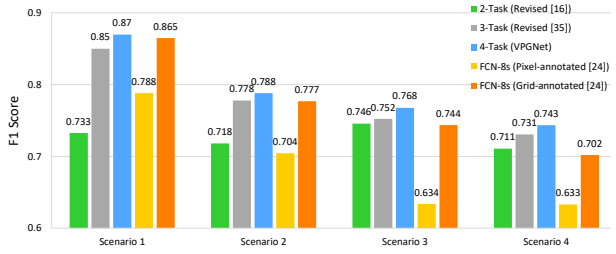


Figure 8. Lane detection score on our benchmark.

riched training and leads to better performance.

In order to see what happens if the VP does not exist, we conducted an additional test on images without the VP (*e.g.* intersection roads or occlusions). Table 4 shows the results of the experiment, demonstrating that the enhancement of feature representation through the VPP task helps to find lanes even when there is no VP. Selected results are shown in the supplementary material.

For the simultaneous detection and classification of lane classes, due to the class imbalance, we measure the F1 score of the top four lane classes by the number of instances. The selected classes are: single white, dashed white, single yellow, and double yellow lines. Table 5 shows the performance of the 3- and 4-Task networks. Except for "no rain, daytime condition", recognition of the single white line is highly improved. This shows that using the VPP task on rainy and night conditions improves the activation of roadway boundaries which are usually marked with single white lines.

5.5. Road Marking Detection and Recognition

In the case of road marking classes, we evaluate the simultaneous detection and classification performance. Due to the dataset imbalance of road marking classes, we measure the recall score of the top four road marking classes by the number of instances. The selected classes are as follows: stop line, straight arrow, crosswalk, and safety zone. Table 6 shows the performance of 3- and 4-Task networks. Except for the stop line class in "no rain, daytime condition", the evaluation results are highly improved. This makes sense because the stop line has horizontal edges which are not

Tab 表4. 无VP集上的车道检测得分（红色：最佳）。

	FCN-8s (pixel)	FCN-8s (grid)	3-Task (revised [35])	4-Task (VPGNet)
No-VP set	0.3310	0.4496	0.4535	0.5234

Tab 表5. 车道类别的同时检测和分类F1评分（红色：最佳） re for

Lane class		Single white	Dashed white	Single yellow	Double yellow
Scenario 1	3-Task	0.55	0.77	0.57	0.32
	4-Task	0.49	0.76	0.58	0.36
Scenario 2	3-Task	0.45	0.67	0.64	0.62
	4-Task	0.52	0.66	0.65	0.61
Scenario 3	3-Task	0.31	0.72	0.70	0.37
	4-Task	0.42	0.73	0.71	0.40
Scenario 4	3-Task	0.27	0.68	0.48	0.36
	4-Task	0.42	0.69	0.42	0.40

Table 6. Simultaneous detection and classification recall score for road marking classes (**Red**: Best).

Road marking class		Stop line	Straight arrow	Crosswalk	Safety zone
Scenario 1	3-Task	0.83	0.46	0.88	0.59
	4-Task	0.78	0.80	0.94	0.80
Scenario 2	3-Task	0.60	0.41	0.81	0.47
	4-Task	0.73	0.65	0.85	0.65
Scenario 3	3-Task	0.33	0.39	0.84	0.47
	4-Task	0.56	0.63	0.93	0.61
Scenario 4	3-Task	0.60	0.48	0.82	0.37
	4-Task	0.80	0.68	0.89	0.38

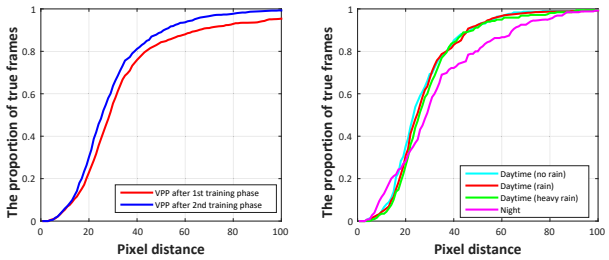


Figure 9. Evaluation on the VPP task.

closely related to the VPP task. Other road markings have shapes that give directions to VP from a geometric perspective. Consequently, responses to those classes become highly activated.

5.6. Vanishing Point Prediction

In the case of a VP, we compare the VPP-only and 4-Task networks. In this manner, we can observe how the VPP is influenced by the lane and road marking detection. Moreover, we compare the performances of each scenario. Figure 9 shows the experimental results. The left graph shows a comparison between two outputs: a prediction after the first phase and a prediction after the second phase. The prediction after the second phase is highly improved meaning that the VPP task gets help from lane and road marking detection tasks. The right graph shows the results of the prediction after the second phase for each scenario.

6. Conclusions

this work, we introduced lane and road marking benchmark that covers four scenarios: daytime (no rain, rain, heavy rain) and night conditions. We have also proposed a multi-task network for simultaneous detection and classification of lane and road markings, guided by a VP. The evaluation shows that the VPGNet model is robust under different weather conditions and performs in real-time. Furthermore, we have concluded that the VPP task enhances both lane and road marking detection and classification by enhancing activation of lane and road markings and the boundary of the roadway.

Acknowledgement

This work was supported by DMC R&D Center of Samsung Electronics Co.

References

[1] M. Aly. Real time detection of lane markers in urban streets. In *IV*, 2008.

[2] O. Bailo, S. Lee, F. Rameau, J. S. Yoon, and I. S. Kweon. Robust road marking detection and recognition using density-based grouping and machine learning techniques. In *WACV*, 2017.

[3] M. Bertozzi and A. Broggi. Real-time lane and obstacle detection on the system. *IV*, 1996.

[4] A. Borji. Vanishing point detection with convolutional neural networks. *arXiv preprint arXiv:1609.00967*, 2016.

[5] A. Borkar, M. Hayes, and M. T. Smith. A novel lane detection system with efficient ground truth generation. *IEEE Transactions on Intelligent Transportation Systems (TITS)*, 13(1):365–374, 2012.

[6] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma. Pcanet: A simple deep learning baseline for image classification? *IEEE Transactions on Image Processing (TIP)*, 24(12):5017–5032, 2015.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[9] H. Deusch, J. Wiest, S. Reuter, M. Szczot, M. Konrad, and K. Dietmayer. A random finite set approach to multiple lane detection. In *ITSC*, 2012.

[10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010.

[11] R. Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[13] J. Greenhalgh and M. Mirmehdi. Automatic detection and recognition of symbols and text on the road surface. In *ICPRAM*, 2015.

[14] B. He, R. Ai, Y. Yan, and X. Lang. Accurate and robust lane detection based on dual-view convolutional neutral network. In *IV*, 2016.

[15] J. Hur, S.-N. Kang, and S.-W. Seo. Multi-lane detection in urban driving environments using conditional random fields. In *IV*, 2013.

[16] B. Huval, T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, P. Rajpurkar, T. Migimatsu, R. Cheng-Yue, et al. An empirical evaluation of deep learning on highway driving. *arXiv preprint arXiv:1504.01716*, 2015.

[17] H. Jung, J. Min, and J. Kim. An efficient lane detection algorithm for lane departure detection. In *IV*, 2013.

[18] J. Kim and M. Lee. Robust lane detection based on convolutional neural network and random sample consensus. In *ICONIP*, 2014.

[19] M. Land, J. Horwood, et al. Which parts of the road guide steering? *Nature*, 377(6547):339–340, 1995.

[20] M. F. Land and D. N. Lee. Where do we look when we steer. *Nature*, 369(6483):742–744, 1994.

[21] J. Li, X. Mei, and D. Prokhorov. Deep neural network for structural prediction and lane detection in traffic scene. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, PP(99):1–14, 2016.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. Ssd: Single shot multibox detector. *arXiv preprint arXiv:1512.02325*, 2015.

[24] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[25] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640*, 2015.

[27] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[28] D. D. Salvucci and R. Gray. A two-point visual control model of steering. *Perception*, 33(10):1233–1248, 2004.

[29] R. Satzoda and M. Trivedi. Vision-based lane analysis: Exploration of issues and approaches for embedded realization. In *CVPR Workshops*, 2013.

[30] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[31] H. Tan, Y. Zhou, Y. Zhu, D. Yao, and K. Li. A novel curve lane detection based on improved river flow and ransa. In *ITSC*, 2014.

[32] D. G. Viswanathan. Features from accelerated segment test (fast), 2009.

[33] P.-C. Wu, C.-Y. Chang, and C. H. Lin. Lane-mark extraction for automobiles under complex conditions. *Pattern Recognition*, 47(8):2756–2767, 2014.

[34] T. Wu and A. Ranganathan. A practical system for road marking detection and recognition. In *IV*, 2012.

[35] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu. Traffic-sign detection and classification in the wild. In *CVPR*, 2016.