Homography Guided Temporal Fusion for Road Line and Marking Segmentation

Shan Wang[1,2] Chuong Nguyen[1] Jiawei Liu[2] Kaihao Zhang[2] Wenhan Luo[3]
Yanhao Zhang[2] Sundaram Muthu[1] Fahira Afzal Maken[1] Hongdong Li[2]
[1]Data61, CSIRO [2]Australian National University [3]Sun Yat-sen University

Abstract

Reliable segmentation of road lines and markings is critical to autonomous driving. Our work is motivated by the observations that road lines and markings are (1) frequently occluded in the presence of moving vehicles, shadow, and glare and (2) highly structured with low intra-class shape variance and overall high appearance consistency. To solve these issues, we propose a Homography Guided Fusion (HomoFusion) module to exploit temporally-adjacent video frames for complementary cues facilitating the correct classification of the partially occluded road lines or markings. To reduce computational complexity, a novel surface normal estimator is proposed to establish spatial correspondences between the sampled frames, allowing the HomoFusion module to perform a pixel-to-pixel attention mechanism in updating the representation of the occluded road lines or markings. Experiments on ApolloScape, a large-scale lane mark segmentation dataset, and ApolloScape Night with artificial simulated night-time road conditions, demonstrate that our method outperforms other existing SOTA lane mark segmentation models with less than 9% of their parameters and computational complexity. We show that exploiting available camera intrinsic data and ground plane assumption for cross-frame correspondence can lead to a light-weight network with significantly improved performances in speed and accuracy. We also prove the versatility of our HomoFusion approach by applying it to the problem of water puddle segmentation and achieving SOTA performance [1].

1. Introduction

Lane mark segmentation aims to achieve pixel-wise classification of road lines and markings simultaneously. Known for its quintessential importance in autonomous driving, lane mark segmentation is also an effective tool for constructing accurate High Definition (HD) maps. Existing works solve individual sub-tasks, *e.g.* drivable area segmentation [33, 9] and lane detection [19, 35], while few address the lane mark segmentation in its entirety [53, 16].

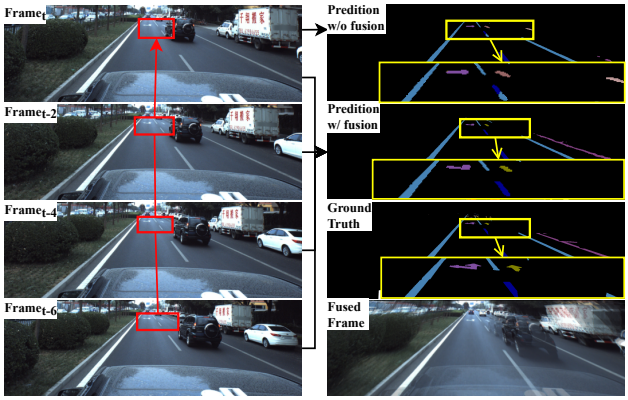


Figure 1: Illustration of the effect of the proposed HomoFusion module that explores the adjacent frames for cues, facilitating the correct classification of (1) a "Straight Arrow", which, with its bottom half occluded by a vehicle, is mistakenly classified as a "Right Turn & Straight Arrow" without the HomoFusion module, and (2) a partially occluded "Dotted Line", which is incorrectly classified as a "Solid Line" without the HomoFusion module. The fused frame in the 4^{th} row and the 2^{nd} column demonstrates the recovered road lines and markings after projecting the previous frames onto the current frame with the estimated homography matrices. The yellow box enlarges the area where mistake classifications are corrected. The red box indicates the spatially corresponding area across the frames. Best viewed in color.

Despite the tremendous progresses in semantic/scene segmentation [52, 42, 26], little attention has been paid to the lane mark segmentation task [53, 16]. Yin *et al*. [53] leverages additional LiDAR information, merging the segmented visual information with the point clouds, to achieve lane mark segmentation. IntRA-KD [16] applies the knowledge distillation technique to improve the efficiency of the lane mark segmentation model.

A major challenge in the lane mark segmentation task is

[1]Code is available at https://github.com/ShanWang-Shan/HomoFusion.

partial occlusion of the road lines and markings caused by the vehicle and surrounding environment leading to false classifications. For example, a partially occluded straight arrow can be easily mistaken as a right turn & straight arrow, as depicted in Fig. 1. However, existing lane mark segmentation methods have not yet made use of the following observations: (1) occlusion of the road lines and markings (due to nearby moving vehicles, shadow, and glare) can be reduced by cross-frame consistency; (2) road lines and markings are highly structured, maintaining high consistency in intra-class shapes as well as in overall appearance, alleviating the requirements on learning complicated features to distinguish different classes with high intra-class variance as in semantic/scene segmentation.

To find complementary information from adjacent frames, we use the ground plane assumption for the road region immediately in front of the camera. Our Homography Guided Fusion (HomoFusion) module achieves temporally consistent representation, recovers partially occluded road lines or markings, and leads to correct classification. The module's success depends on accurate cross-frame spatial correspondence, which we achieve using a homography transformation matrix estimated with available intrinsic and extrinsic parameters of the on-vehicle camera, and the road surface normal estimated by our novel optimization method called Road Surface Normal Estimator (RSNE).

Using the highly structured nature of road lines and markings could reduce the complexity of the detection problem and make it suitable to run on edge devices. To this end, we employ (1) a lightweight encoder to represent the visual information in feature spaces, and (2) a cross-frame pixel-to-pixel attention mechanism in our HomoFusion module, instead of a more computationally expensive global attention mechanism employed by other methods. These design decisions allow the proposed model to outperform the existing lane mark segmentation models with less than 9% of their parameters and Giga Floating Point Operations (GFLOPs) We summarize our contributions as follows:

- We propose a HomoFusion module that uses ground plane assumption and the adjacent frames for temporally consistent representations for accurate classification of partially occluded road lines and markings.
- We design a novel estimator RSNE for road surface normal, which, combined with camera intrinsic and extrinsic parameters readily available on an autonomous vehicle, yields accurate homography matrix between frame pairs. RSNE simplifies the 8 degrees of freedom (DoF) of homography problem to a 2 DoF of normal vector problem.
- We present a lightweight lane mark segmentation model that achieves better performance than the state-of-the-art (SOTA) methods with significantly reduced model complexity and computation requirements.

2. Related Work

2.1. Lane Detection

Traditional lane detection approaches rely on handcrafted features such as color [44, 51], edge [23] and texture [25], which are limited in complex scenarios [48, 7]. Recent advances in deep neural networks have led to significant performance improvements, with methods like message-passing networks [36] and attention-based modules [45]. [16, 37, 38, 54, 27] formulate lane detection as a row-wise classification task based on grid division of the input image. PolyLaneNet [46] is the first parametric prediction method, which outputs polynomials to represent each lane. BEVFormer [24] and PETRv2 [28] employ transformer mechanisms to streamline the conversion of perspective views into bird's-eye views, thereby eliminating variations in object size caused by perspective. Although the above methods achieve impressive performance, all of them focus on detecting the lane lines and ignore the basic road elements like arrow signal in Fig. 1. Furthermore, lane detection differs from segmentation tasks in that it focuses on identifying the boundaries of the drivable area, rather than the real shape of the lanes.

2.2. Road Line and Marking Segmentation

The identification of road elements is crucial for ensuring safety in autonomous driving systems, but there are few works addressing this problem. Hou *et al.* [16] introduced a distillation approach that demonstrates competitive performance in lane mark segmentation. Yin *et al.* [53] used an LSTM-based network to segment images with DeepLabv3+ [8] and then merged them with point clouds to assist with lane segmentation. Unfortunately, these methods do not address the challenges of occlusion, shadows, and glare that frequently occur in real-world driving scenarios. Recently, Zhang *et al.* [56] introduced global memory information from previous frames to enhance local information for video lane segmentation. However, this method introduces attention to the entire image, making it memory-intensive and inefficient for autonomous driving scenarios.

2.3. Homography Estimation

Homography estimation methods can be divided into non-deep and deep learning-based approaches. Non-deep methods estimate the homography using feature extraction, feature matching, and outlier rejection. SIFT [29], SURF [5, 4], ORB [39], LPPM [30], GMS [6], and BEBLID [41] are commonly used for feature extraction, and RANSAC [12], MAGAC [2], and LRLS [15] are applied for outlier rejection. Recently, many deep learning-based approaches

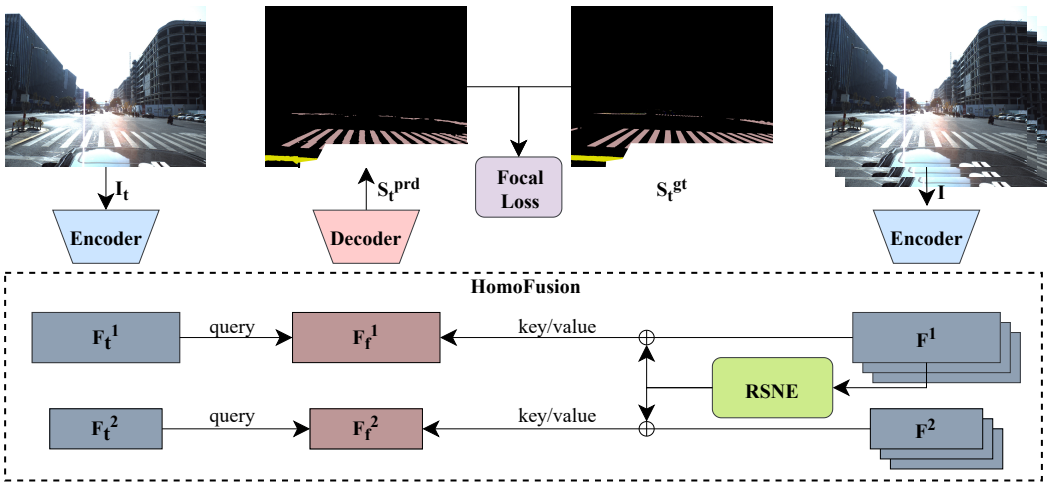


Figure 2: Overview of our proposed model consisting of a pair of lightweight encoder and decoder, our proposed HomoFusion module, and our proposed Road Surface Normal Estimator (RSNE). A sequence of frames $\mathbf{I}$, including a target frame $\mathbf{I_t}$ and $n-1$ previous frames, are encoded into the feature representations ($\mathbf{F}^l$). RNSE estimates the road surface normal vector, which, combined with the camera intrinsic and extrinsic parameters, yields a homography matrix between each frame pair, establishing cross-frame spatial correspondences. HomoFusion uses pixel-to-pixel attention mechanism to obtain temporally consistent representation for on-road pixels of the current frame with the spatial correspondence across frames as guidance. Finally, the decoder decodes and upsamples the temporally consistent feature representations to produce the lane mark segmentation prediction ($\mathbf{S_t^{prd}}$).

have been proposed, such as DeTone *et al.* [10]'s VGG-like architecture, Nowruzi *et al.* [11], Le *et al.* [20]'s and Man *et al.* [31]'s cascaded VGG-like networks, and Chang *et al.* and Zhao *et al.*'s incorporation of the Lucas-Kanade (LK) algorithm with deep networks. However, these methods do not utilize known camera intrinsic and extrinsic parameters [2] across frames for homography matrix estimation, which can effectively reduce the search dimensions.

3. Method

3.1. Overview

As illustrated in Fig. 2, our proposed framework takes a sequence of frames, denoted as $\mathbf{I} = \{\mathbf{I}_i | i = \{t-(s-1)\Delta t\}_{s=1}^{n}\}$, as input. This sequence consists of a current frame $\mathbf{I}_t$ and $n-1$ previous frames sampled at a fixed time interval Δt. The output of the framework is a lane mark segmentation map $\mathbf{S}_t$ for the current frame. In addition, the framework includes a novel HomoFusion module, which uses the homography transformation guided by the proposed RSNE to fuse the feature map of the current frame with those of the previous frames. We denote the encoded feature representation for the t^{th} frame as $\mathbf{F}_t = \{\mathbf{F}_t^l \in \mathbb{R}^{H_l \times W_l \times C_l}\}_{l=1}^{L}$ where l indicates the level of the feature map. The shallowest and deepest levels are represented by 1 and L, respectively. H_l, W_l, and C_l represent the height, width, and channel number of feature maps in level l. We denote all encoded features as $\mathbf{F} = \{\mathbf{F}_i^l | i = \{t-(s-1)\Delta t\}_{s=1}^{n}\}$. The proposed HomoFusion module, detailed in Sec. 3.2, uses a cross-frame pixel-to-pixel attention mechanism to fuse the feature map ($\mathbf{F}_f = \{\mathbf{F}_f^l\}_{l=1}^{L}$) of the current frame with those of the previous frames. The homography transformation matrix between the target frame and each of the previous frames is calculated using the estimated normal vector of the road surface $\mathbf{n}$ obtained by the proposed RSNE, as described in Sec. 3.3. Finally, the decoder produces the segmentation prediction ($\mathbf{S}_t$) for the target frame based on the fused feature map.

[2] In our experiments, both camera intrinsic and extrinsic parameters are provided by the dataset. In real-world scenarios, these parameters can be obtained through camera calibration and autonomous vehicle pose obtained from the pose estimation framework, which is already known in the in-car system and free of cost.

3.2. Homography Guided Fusion (HomoFusion)

Our proposed HomoFusion module employs a pixel-to-pixel attention mechanism to fuse spatially corresponding pixels across frames. This mechanism is achieved by projecting the pixels of the current frame onto the previous frames through an accurate homography transformation. Since road lines and markings are strictly on-road, the search area can be limited to the road surface, which is mostly a plane, at least within the immediate front of the vehicle where markings are readable. Therefore, only the homography transformation is needed to accurately map road pixels between frames. According to the standard inverse homography [14, 57, 49], for each on-road pixel $p_t =$

Figure 3: Illustration of sample points. (Right) Sample points in the current frame. (Left/Middle) Correspondence of a sample point in previous frames. The red point coordinate is calculated by using the correct normal, while the cyan point coordinate is calculated by using the initial (incorrect) normal.

$(u, v)^\top$ of the current frame, its spatially corresponding pixels of the reference frames $\{p_i | i = \{t-(s-1)\Delta t\}_{s=1}^{n}\}$ can be computed using Eq. (1).

$$p_i \propto \mathbf{K}(\mathbf{R}_i - \frac{\mathbf{t}_i \mathbf{n}^\top}{d})\mathbf{K}^{-1}(p_t \oplus 1), \tag{1}$$

where $\propto$ represents proportional, $\mathbf{K}$ is the intrinsic matrix of the on-vehicle camera, $\mathbf{R}_i$ and $\mathbf{t}_i$ represent the relative rotation and translation between the current frame $\mathbf{I}_t$ and the i^{th} previous frame $\mathbf{I}_i$. $\mathbf{n}$ is the normal of the observed road surface, which is estimated in Sec. 3.3. d is the vertical distance between the on-vehicle camera and the road surface obtained from the camera calibration. $\oplus 1$ converts the point coordinates to homogeneous coordinates. The on-road pixels of the current frame are sampled from a triangular area in front of the car, as illustrated in Fig. 3. Figure 3 also demonstrates that the shape and category of a left-turn can be recovered by exploring the spatially corresponding area in the previous frames under the guidance of an accurate homography transformation.

With the spatial correspondence across frames, the proposed HomoFusion module computes a temporally consistent representation for each sampled pixel of the current frame by employing temporal attention mechanism on the feature presentations of spatially corresponding pixels. The mechanism assigns the pixel representation of current frame ($\mathbf{F}_t[p_t]$) as a query and all frames ($\{\mathbf{F}_i[p_i]\}$) as keys and values. In contrast to the existing attention mechanism [50], we propose to (1) omit the spatial encoding in the presence of obtained spatial correspondence across frames, and (2) apply an L2 normalization on the query and the keys to obtain more robust (deep) feature similarities (Eq. 2) across various lighting environments.

$$\mathbf{a}_i = \frac{\mathbf{F}_t[p_t]}{\|\mathbf{F}_t[p_t]\|_2} \cdot \frac{\mathbf{F}_i[p_i]}{\|\mathbf{F}_i[p_i]\|_2}, \tag{2}$$

where $\|.\|_2$ indicates an L_2 normalization. The similarity is normalized by softmax function as given in Eq. 3:

$$\mathbf{W}_i = \frac{\exp \mathbf{a}_i}{\sum_i \exp \mathbf{a}_i}, \tag{3}$$

before being used as weight to fuse the spatially corresponding pixels across the frames as shown in Eq. 4:

$$\mathbf{F}_f[p_t] = \mathbf{F}_t[p_t] + \sum_i \mathbf{W}_i \mathbf{F}_i[p_i]. \tag{4}$$

3.3. Road Surface Normal Estimator (RSNE)

Accurate estimation of road surface normal is crucial for establishing spatial correspondence across frames. Generally, on-vehicle cameras have a fixed angle with respect to the ground level. However, the actual road surface normal can vary due to the unevenness of the road surface, *e.g.* uphill or downhill roads, and sloped turns, *etc*. Assuming that the initial road surface normal is a vector being perpendicular to a horizontal plane and pointing upwards, we propose to obtain an accurate road surface normal with an optimization process that repeats for each new coming frame. The proposed RSNE iteratively updates the estimated road surface normal with the projection error of on-road pixels from the current frame onto the previous frames. Specifically, we sample m pixels $\{p_t^j\}_{j=1}^m$ from the current frame that are likely to be on-road, as illustrated in Fig. 3, and compute the spatially corresponding pixels $\{p_i^j\}_{j=1}^m$ in the previous frames using the homography matrix computed with the estimated road surface normal as defined in Eq. 1. The estimated road surface normal is optimized in an iterative manner through the following procedure: (a) computing a residual between the low-level features of spatially corresponding pixels across the frames with Eq. 5

$$\mathbf{r}_i^j = \mathbf{F}_i^1[p_i^j] - \mathbf{F}_t^1[p_t^j] \in \mathbb{R}^c, \tag{5}$$

where the residual is set to 0 for the sample point whose projections in previous frames fall outside the frame area, and computing the overall error with Eq. 6

$$\mathbf{E} = \sum_{i,j} \rho(\|\mathbf{r}_i^j\|_2^2), \tag{6}$$

where $\|.\|_2^2$ shows squared norm and $\rho()$ is a robust cost function proposed by Barron [3]. This nonlinear least-square cost is iteratively minimized towards correct normal estimation using the Levenberg–Marquardt algorithm [21, 32]; (b) computing a new road surface normal value by adding the update given by Eq. 12.

The road surface normal vector is a unit vector with two degrees of freedom (DoF), which can be decomposed into a pitch angle θ and a roll angle ϕ [3], as shown in Eq. 7.

$$\mathbf{n} = (-\sin\phi\cos\theta,\ -\cos\phi\cos\theta,\ \sin\theta)^\top. \tag{7}$$

The Jacobian of the residual function w.r.t. the pitch angle and the roll angle is defined in Eq. 8 by applying the chain

[3] An explanation of the pitch angle θ and roll angle ϕ is provided in the appendix.

rule,

$$\mathbf{J}_i^j = \frac{\partial \mathbf{r}_i^j}{\partial \delta} = \frac{\partial \mathbf{F}_i^1[p_i^j]}{\partial p_i^j} \frac{\partial p_i^j}{\partial \mathbf{n}} \frac{\partial \mathbf{n}}{\partial \delta}, \tag{8}$$

where δ represents the to-be-optimized target (θ or ϕ), and $\frac{\partial \mathbf{F}_i^1[p_i^j]}{\partial p_i^j}$ is gradient from 2D interpolation. $\frac{\partial p_i^j}{\partial \mathbf{n}}$ is the Jacobian of the homography transformation (Eq. (1)) w.r.t. $\mathbf{n}$, as defined in Eq. 9

$$\frac{\partial p_i^j}{\partial \mathbf{n}} = -\frac{1}{d} \mathbf{K} \mathbf{t}_i (\mathbf{K}^{-1}(p_t^j \oplus 1))^\top, \tag{9}$$

and $\frac{\partial \mathbf{n}}{\partial \delta}$, defined in Eq. 10 and Eq. 11, is the Jacobian of Eq. (7) w.r.t. θ and ϕ respectively

$$\frac{\partial \mathbf{n}}{\partial \theta} = (-\frac{\mathbf{n}_1 \mathbf{n}_3}{\sqrt{1-\mathbf{n}_3^2}},\ -\frac{\mathbf{n}_2 \mathbf{n}_3}{\sqrt{1-\mathbf{n}_3^2}},\ \sqrt{1-\mathbf{n}_3^2}), \tag{10}$$

$$\frac{\partial \mathbf{n}}{\partial \phi} = (\mathbf{n}_2,\ -\mathbf{n}_1,\ 0), \tag{11}$$

where $\mathbf{n}_k$ represents the k-th element of $\mathbf{n}$. The detailed derivation is included in appendix (Sec. 1). We obtain Hessian matrices as $\mathbb{H} = \mathbf{J}^\top \rho' \mathbf{J}$, where ρ' is the derivative of function $\rho()$. We compute the update by damping the Hessian and solving the linear system, as shown in

$$\triangle \delta = -(\mathbb{H} + \lambda \operatorname{diag}(\mathbb{H}))^{-1} \mathbf{J}^\top \rho' \mathbf{r}, \tag{12}$$

where λ is the damping factor [40], balancing between the Gauss-Newton ($\lambda = 0$) and gradient descent ($\lambda = \infty$). Our entire optimization process is differentiable.

The algorithm is illustrated in Alg. 1. The normal vector $\mathbf{n}$ uses low-level features $\mathbf{F}^1$. While the high-level features $\mathbf{F}^2$ focus on segmentation feature expression, the low-level features $\mathbf{F}^1$ are supervised to extract useful features for both segmentation and road surface normal estimation. As both tasks focus on on-road objects, they can benefit each other.

Algorithm 1 Normal Optimization

Input:
The 1-st Level Feature Maps: $\mathbf{F}^1 = \{\mathbf{F}_i^1\}$;
Sample Points Coordinates: $p_t = \{p_t^j\}_{j=1}^m$;
Initial Pitch and Roll Value: $(\theta_0,\ \phi_0) = (0.15,\ 0)$;
Damping Factor: $\lambda = (\lambda_\theta\ \lambda_\phi)$;
Hyper-parameter:
$\mathbf{k}$; ▷ Maximum loop count, empirically set to 20
α; ▷ Convergence threshold, empirically set to 0.0001
Output: Optimized Pitch and Roll $(\theta,\ \phi)$;

function OPT($\theta_0, \phi_0, \mathbf{F}^1, p_t, \lambda$)
  Derive point features $\mathbf{F}_t^1[p_t]$ from $\mathbf{F}_t^1$;
  for $k \leftarrow 1$ to **k do**
    Calculate point coordinates p_i in $\mathbf{F}_i^1$ (Eq. 1);
    Derive point features $\mathbf{F}_i^1[p_i]$ from $\mathbf{F}_i^1$;
    Calculate residual $\mathbf{r}_i$ (Eq. 5);
    Calculate observe error $\mathbf{E}$ (Eq. 6);
    Calculate robust cost $\rho(\mathbf{E})$ and its derivation ρ';
    Construct $\mathbf{J}$ and its Hessian matrices $\mathbb{H}$ (Eq. 8);
    Obtain $\triangle\delta$ by Cholesky decomposition (Eq. 12);
    Update Normal as $(\theta,\ \phi)_k \leftarrow (\theta,\ \phi)_{k-1} + \triangle\delta$;
    if MAX($\triangle\delta$) $< \alpha$ **then**
      Break;
    end if
  end for
end function

4. Experiments

4.1. Implementation Details

We use a pre-trained EfficientNet [47] to extract image features at two different scales, $4\times$ and $16\times$ down-scaling, with 64 and 128 channels, respectively. To focus on the bottom region of the input image, where road marks are typically visible, we crop only the bottom 40% of the input image, following the same approach in the prior work [16]. The processed image size is set to 272×848. Our decoder consists of bi-linear up-sampling layers and convolution layers, which up-sample the high-level features and increase the resolution by a factor of 4 at each level. We trained our model using the AdamW optimizer [55] for 30,000 iterations on two NVIDIA RTX 3090 GPUs, with a learning rate of 4×10^{-3}.

4.2. Datasets

We conduct our experiments on two datasets, ApolloScape [17] and ApolloScape Night. Other datasets were deemed unsuitable due to the absence of camera extrinsics, as observed in SDLane [18] and VLI-100 [56], inadequate frame overlap, or the lack of diverse segmentation labels, exemplified by Waymo Open Dataset [43]. Further elaboration on these reasons can be found in the appendix.

ApolloScape. The ApolloScape dataset is a large-scale dataset that can be used for localization and segmentation tasks. It consists of 38 distinct classes and poses various challenges including occlusions and tiny road markings. To provide accurate camera poses, the vehicle is equipped with customer-grade GPS/IMU. However, for the lane mark segmentation task, only $41,201$ annotated images out of more than $110,000+$ are associated with camera pose information. We used these $41,201$ images for our experiments since our approach relies on camera poses. We divided them into $35,173$ training images and $6,028$ validation images. Because our approach uses sequential information, we select the validation set with completely different trajectories to ensure a fair comparison.

ApolloScape Night. Since there is a lack of datasets for

Table 1: Comparison with SOTA Methods on the ApolloScape [17] and Apolloscape Night Datasets and Running on an NVIDIA RTX 3090 GPU. "18 mIoU" Represents the Mean IoU of 18 Types of Lane Markers Selected by ApolloScape Official Metrics. "36 mIoU" Represents the Mean IoU of All Unignorable 36 Types.

Methods	Frame Count	Backbone	Params (M)↓	GFLOPs↓	FPS (f/s)↑	ApolloScape [17]		ApolloScape Night	
						18 mIoU↑	36 mIoU↑	18 mIoU↑	36 mIoU↑
IntRA-KD[16]	1	ResNet-101	65.6	5159.4	10.8	42.1	24.6	29.8	16.7
SegFormer[52]	1	MiT-B1	13.5	1048.8	**43.8**	52.3	32.1	38.3	23.1
CFFM [42]	4	MiT-B1	15.3	1192.6	22.7	53.2	32.7	39.1	23.6
MMA-Net [56]	4	ResNet-50	57.9	723.2	20.6	52.9	31.4	38.8	23.2
HomoFusion(ours)	4	EfficientNet-B6	**1.24**	**61.2**	25.4	**59.3**	**35.9**	**44.8**	**26.6**

Figure 4: Sample images from the ApolloScape Night dataset. Top: original daytime images from the ApolloScape dataset. Bottom: synthesized night-time images.

night-time lane mark segmentations, we created an artificial dataset called ApolloScape Night from the ApolloScape dataset using a cross-domain generation network [1]. This allows us to evaluate our proposed model on a challenging dataset with poor lighting conditions, road reflection, and glare. Fig. 4 displays some sample images from the dataset.

4.3. Evaluation Metrics

In accordance with the guidelines of the ApolloScape benchmark [17], we used mean intersection-over-union (mIoU) as the evaluation criterion. ApolloScape contains 38 different types of lane markers, including two ignorable labels (noise and ignored). Of these labels, 18 categories are used in the official evaluation metrics. To provide a comprehensive evaluation of our approach, we report the evaluated mIoU results for both the selected 18 categories, denoted as "18 mIoU", and all 36 categories, denoted as"36 mIoU". By reporting both sets of results, we provide a more complete picture of the performance of our method on the ApolloScape and ApolloScape Night datasets.

4.4. Comparison with State-of-the-art Methods

To evaluate the performance of our proposed method, we compared it with SOTA algorithms, including (a) IntRA-KD [16], (b) SegFormer [52], (c) CFFM [42], and (d) MMA-Net [56], on the ApolloScape [17] and ApolloScape Night datasets. To ensure a fair comparison, we retrained each model on the same subset of the training set with the same input resolution.

The results of the comparison are presented in Table 1. Our approach outperforms the existing SOTA methods on both the ApolloScape [17] and ApolloScape Night datasets while having less than 9% of their parameters or computational overhead.

In addition, we compared our proposed method with CFFM [42] which also uses adjacent frames to enhance the representation of the current frame. Our method is more efficient in terms of model complexity, with a complexity of $\mathcal{O}(HWC)$ for feature extraction and $\mathcal{O}(HWC^2)$ for cross-frame feature fusion. In contrast, the complexity of CFFM is $\mathcal{O}(H^2W^2C) + \mathcal{O}(HWC^2)$ for feature extraction and $\mathcal{O}(HWEC) + \mathcal{O}(HWC^2)$ for cross-frame feature fusion, where E is calculated by their receptive field and pooling kernel size. This explains why our method has a lower computational requirement for the cross-frame pixel-to-pixel attention mechanism compared to the global attention mechanism used by CFFM.

The qualitative results of both the existing SOTA models and our proposed model are shown in Fig. 5. The results indicate that our approach achieves superior segmentation performance on road lines and markings, even under adverse conditions such as occlusion, road reflection, and poor lighting. Additional visualization examples for various categories can be found in the appendix.

4.5. Study on Hyper-parameters

Number of Frames (n). We conducted a study on the impact of the number of frames while setting the sampling frame gap $\triangle t = 2$. The results presented in Fig. 6 (a)-(b) demonstrate that the performance of our method improves on both ApolloScape and ApolloScape Night datasets with an increase in the number of explored adjacent frames. However, the performance improvement becomes negligible when the number of frames exceeds 4. Considering the monotonically increasing relationship between the model complexity and the number of frames, we set $n = 4$ to strike a balance between model performance and complexity.

Sampling Frame Gap ($\triangle t$). In the study on the effect of the sampling frame gap, we set the number of frames

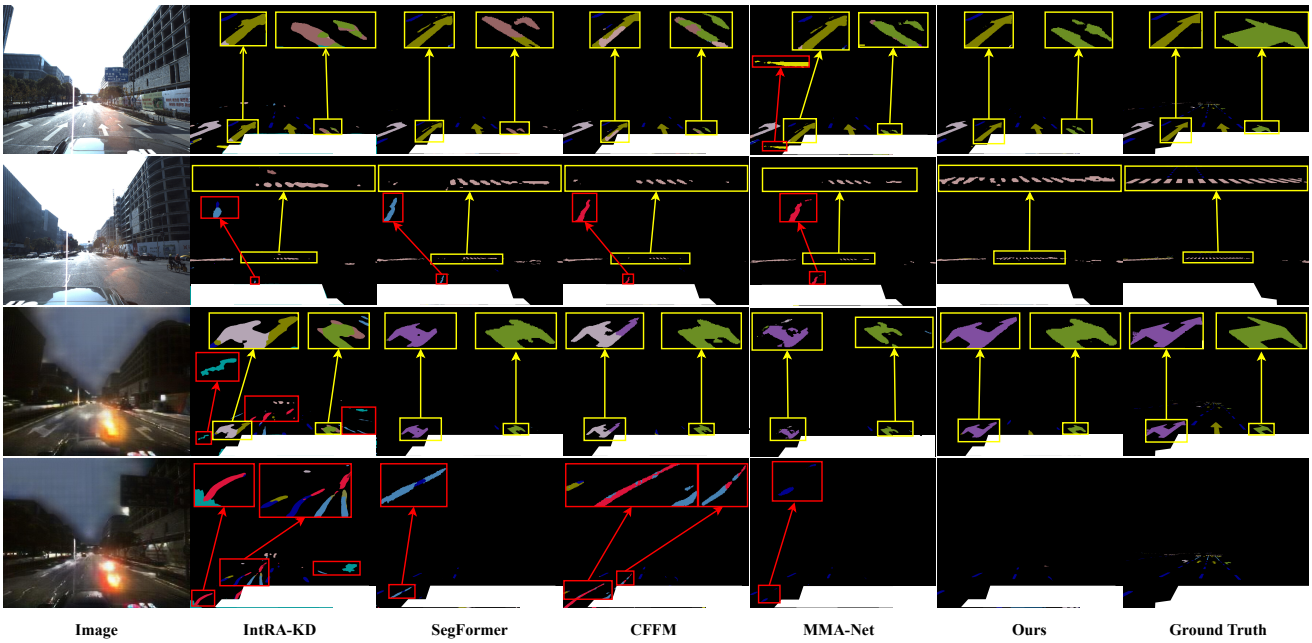
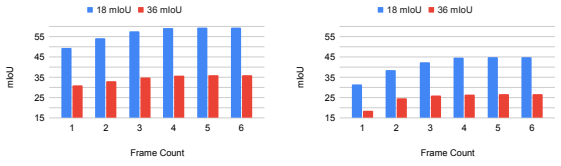
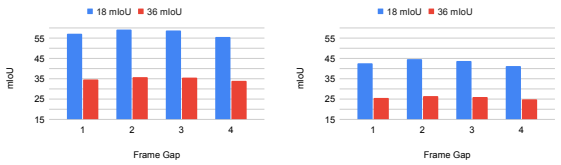


Figure 5: Qualitative comparison with SOTA methods. The top two examples are from the ApolloScape [17] dataset, and the bottom two examples are from the ApolloScape Night dataset. Yellow boxes highlight the area of interest for better visualization. Red boxes indicate false-positive segmentation predictions. Best viewed in color.



(a) Number of frames (AS))

(b) Number of frames (ASN)



(c) Sampling frame gap (AS)

(d) Sampling frame gap (ASN)

Figure 6: The impact of the number of frames n on ApolloScape (AS) and ApolloScape Night (ASN) datasets is shown in (a) and (b), while (c) and (d) show the effect of the sampling frame gap $\triangle t$.

$n = 4$. Fig. 6 (c)-(d) indicates that the optimal performance is achieved at $\triangle t = 2$. This can be attributed to the balance between the sufficient overlap of road area across the frames and a wide enough temporal gap that allows the displacement of the source of occlusion, *e.g.* moving vehicle.

4.6. Discussion

All ablation studies are conducted on the ApolloScape dataset with the same training strategies described in Sec. 4.1. We set the number of frames to $n = 4$ and a sampling frame gap of $\triangle t = 2$.

Table 2: Ablation Study of our Proposed HomoFusion and RSNE on the ApolloScape Dataset

	18 mIoU↑	36 mIoU↑
w/o HomoFusion	49.4	31.0
HomoFusion w/o HomoGuide	51.8	31.3
HomoFusion w/o RSNE	57.4	35.3
HomoFusion w/ RGB RSNE	54.9	33.8
Noisy Extrinsic	57.5	35.1
HomoFusion Full	**59.3**	**35.9**

Impact of HomoFusion. The core component of our approach is the HomoFusion module, which allows the use of complementary information from sequential frames. Tab. 2 demonstrates that our model with the HomoFusion module (4 frames) achieves better performance than the variant without HomoFusion module (1 frame), illustrating the effectiveness of the temporally consistent feature representation obtained with HomoFusion module in addressing the partial occlusion issues. We also study the effect directly associating the pixels across frames based on their coordinates by using identity matrix as the homography matrix, denoted as "w/o HomoGuide". It shows that such an incorrect cross-frame spatial correspondence, although leveraging additional frames, provides limited help in improving the model performance.

Impact of RSNE. We study the impact of the proposed RSNE in the following model variants each comprising the HomoFusion modules: (a) without RSNE ("HomoFusion

w/o RSNE"), where the homography matrix is estimated with the initial road surface normal ($\mathbf{n}_0$), (b) with RSNE using RGB information ("HomoFusion w/ RGB RSNE") and (c) with RSNE with features ("HomoFusion Full"). Results in Tab. 2 show that "HomoFusion Full" achieves better performance than "HomoFusion w/o RSNE", demonstrating the effectiveness of the iterative optimization in estimating a more accurate homography matrix that subsequently provides more accurate guidance on cross-frame spatial correspondences. It can be also observed that "HomoFusion Full" significantly outperforms "HomoFusion w/ RGB RSNE" indicating that deep features are more robust than RGB information in estimating an accurate road surface normal. This can be attributed to the fact the feature representations are less prone to adverse noises caused by environmental factors, *e.g.* brightness. Fig. 7 shows that the fused frame with road surface normal estimated with features has better alignment in road lines and markings than that with RGB information.

(a) Normal estimation with RGB (b) Normal estimation with feature

Figure 7: Fused frame sequence with the homography transformation between the frame pairs. The road surface normal estimated with feature representation enables more accurate homography transformation than with RGB information, resulting in better alignment of road lines and markings across frames.

Robustness against Extrinsic Noise. We investigate the robustness of our method to extrinsic noise by introducing random translation errors of less than 1 meter and rotation errors of up to 30°. The results are shown in "Noisy Extrinsic" of Tab. 2, which demonstrate that our method is capable of mitigating extrinsic noise and remains robust.

Impact of Backbone. In order to ensure a fair comparison with CFFM [4] [42] and MMA-Net [56], we implement our method with 'MiT-B1' and 'ResNet-50' backbone and report its performance in Tab. 3. Our method still outperforms CFFM [42] and MMA-Net [56] significantly. Additionally, we test our method with 'EfficientNet-B4/5' backbones, and the results show that our method has relatively stable performance even when using smaller backbones.

Application to Another Task. Given our method's ability to accurately align road surface objects across adjacent frames, we also try it with the task of detecting water hazards. Specifically, our approach aims to perform binary segmentation to identify water puddles in the road. The previous work on this task T3D-FCN [22] also exploits temporal information, as water puddles can appear differently at different angles and distances as a vehicle moves. We use the Puddle-1000 dataset [13], which includes On-road and Off-road datasets. The frames obtained from an onboard camera on bumpy roads are not stable and contain small extrinsic noise. The intrinsic information of the camera is provided by the dataset, while the global extrinsic information of the camera is obtained using ORB-SLAM [34]. Our experiments are conducted using the current and seven previously consecutive frames with a 240×320 image size, the same settings as T3D-FCN [22]. The results, presented in Tab. 4, demonstrate the high effectiveness of our method. More visual details of this task can be found in the Appendix.

[4]We use a heavy backbone in the competing methods as their publicly available code uses this backbone

Table 3: Comparison of our Proposed HomoFusion with Different Backbones on ApolloScape Dataset

Backbones	Params(M)↓	GFLOPs↓	18 mIoU↑	36 mIoU↑
MiT-B1	13.64	89.8	57.2	35.3
ResNet-50	9.12	189.7	55.3	34.6
EfficientNet-B4	**0.8**	**60.8**	57.0	35.3
EfficientNet-B5	1.02	61.2	57.9	35.2
EfficientNet-B6	1.24	61.2	**59.3**	**35.9**

Table 4: Comparison of water puddle segmentations on Puddle-1000 dataset

Dataset	Methods	F1-meas↑	Prec↑	Rec↑
On-road	FCN-8s-FL-RAU [13]	0.70	0.68	0.72
	T3D-FCN [22]	0.68	0.79	0.62
	HomoFusion (ours)	**0.80**	**0.81**	**0.78**
Off-road	FCN-8s-FL-RAU [13]	0.81	0.87	0.77
	T3D-FCNU [22]	0.73	0.87	0.63
	HomoFusion (ours)	**0.87**	0.87	**0.87**

5. Conclusion

The proposed lightweight lane mark segmentation model presented in this paper offers superior performance with reduced model complexity compared to SOTA approaches. The integration of the HomoFusion module and Road Surface Normal Estimator provides an accurate classification of partially occluded, shadowed, and/or glare-affected road lines and markings by leveraging adjacent frames and pixel-to-pixel attention mechanisms. Moreover, the novel approach of using the road surface normal to guide spatial correspondences across frames has significant implications for a wide range of applications in autonomous driving, including road surface inspection, water, and other hazard detections. However, the current model relies on available camera extrinsics and is somewhat impacted by extrinsic noise. Future work will focus on optimizing both the normal vector and camera extrinsic parameters to improve the performance and make the method more robust to challenging environmental conditions. Overall, the proposed method has the potential to significantly advance the field of autonomous driving and related applications.

6. Acknowledgements

The research is funded in part by an ARC Discovery Grant (grant ID: DP220100800) to HL.

References

[1] Vinicius F Arruda, Thiago M Paixão, Rodrigo F Berriel, Alberto F De Souza, Claudine Badue, Nicu Sebe, and Thiago Oliveira-Santos. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[2] Daniel Barath, Jiri Matas, and Jana Noskova. Magsac: marginalizing sample consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10197–10205, 2019.

[3] Jonathan T. Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[4] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[6] JiaWang Bian, Wen-Yan Lin, Yasuyuki Matsushita, Sai-Kit Yeung, Tan-Dat Nguyen, and Ming-Ming Cheng. Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4181–4190, 2017.

[7] Amol Borkar, Monson Hayes, and Mark T Smith. A novel lane detection system with efficient ground truth generation. *IEEE Transactions on Intelligent Transportation Systems*, 13(1):365–374, 2011.

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[9] Zhe Chen, Jing Zhang, and Dacheng Tao. Progressive lidar adaptation for road detection. *IEEE/CAA Journal of Automatica Sinica*, 6(3):693–702, 2019.

[10] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016.

[11] Farzan Erlik Nowruzi, Robert Laganiere, and Nathalie Japkowicz. Homography estimation from image pairs with hierarchical convolutional networks. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 913–920, 2017.

[12] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[13] Xiaofeng Han, Chuong Nguyen, Shaodi You, and Jianfeng Lu. Single image water hazard detection using fcn with reflection attention units. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 105–120, 2018.

[14] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[15] Paul W Holland and Roy E Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics-theory and Methods*, 6(9):813–827, 1977.

[16] Yuenan Hou, Zheng Ma, Chunxiao Liu, Tak-Wai Hui, and Chen Change Loy. Inter-region affinity distillation for road marking segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12486–12495, 2020.

[17] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 954–960, 2018.

[18] Dongkwon Jin, Wonhui Park, Seong-Gyun Jeong, Heeyeon Kwon, and Chang-Su Kim. Eigenlanes: Data-driven lane descriptors for structurally diverse lanes. In *CVPR*, 2022.

[19] Yeongmin Ko, Younkwan Lee, Shoaib Azam, Farzeen Munir, Moongu Jeon, and Witold Pedrycz. Key points estimation and point instance segmentation approach for lane detection. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):8949–8958, 2022.

[20] Hoang Le, Feng Liu, Shu Zhang, and Aseem Agarwala. Deep homography estimation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7652–7661, 2020.

[21] Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944.

[22] Juntao Li, Chuong Nguyen, and Shaodi You. Temporal 3d fully connected network for water-hazard detection. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–5, 2019.

[23] Yadi Li, Liguo Chen, Haibo Huang, Xiangpeng Li, Wenkui Xu, Liang Zheng, and Jiaqi Huang. Nighttime lane markings recognition based on canny detection and hough transform. In *IEEE International Conference on Real-time Computing and Robotics*, pages 411–415. IEEE, 2016.

[24] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022.

[25] Zuo-Quan Li, Hui-Min Ma, and Zheng-Yu Liu. Road lane detection with gabor filters. In *International Conference on Information System and Artificial Intelligence*, pages 436–440. IEEE, 2016.

[26] Jiawei Liu, Jing Zhang, Yicong Hong, and Nick Barnes. Learning structure-aware semantic segmentation with image-level supervision. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[27] Lizhe Liu, Xiaohao Chen, Siyu Zhu, and Ping Tan. Condlanenet: a top-to-down lane detection framework based on

conditional convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3773–3782, 2021.

[28] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petrv2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*, 2022.

[29] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[30] Jiayi Ma, Ji Zhao, Junjun Jiang, Huabing Zhou, and Xiaojie Guo. Locality preserving matching. *International Journal of Computer Vision*, 127(5):512–531, 2019.

[31] Yunze Man, Xinshuo Weng, Xi Li, and Kris Kitani. Groundnet: Monocular ground plane normal estimation with geometric consistency. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2170–2178, 2019.

[32] Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[33] Annika Meyer, N. Ole Salscheider, Piotr F. Orzechowski, and Christoph Stiller. Deep semantic lane segmentation for mapless driving. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 869–875, 2018.

[34] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[35] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Towards end-to-end lane detection: an instance segmentation approach. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 286–291, 2018.

[36] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Spatial as deep: Spatial cnn for traffic scene understanding. volume 32, 2018.

[37] Jonah Philion. Fastdraw: Addressing the long tail of lane detection by adapting a sequential prediction network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11582–11591, 2019.

[38] Zequn Qin, Huanyu Wang, and Xi Li. Ultra fast structure-aware deep lane detection. In *European Conference on Computer Vision*, pages 276–291. Springer, 2020.

[39] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011.

[40] Paul-Edouard Sarlin, Ajaykumar Unagar, Måns Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, and Torsten Sattler. Back to the Feature: Learning robust camera localization from pixels to pose. In *CVPR*, 2021.

[41] Iago Suárez, Ghesn Sfeir, José M Buenaposada, and Luis Baumela. Beblid: Boosted efficient binary local image descriptor. *Pattern recognition letters*, 133:366–372, 2020.

[42] Guolei Sun, Yun Liu, Henghui Ding, Thomas Probst, and Luc Van Gool. Coarse-to-fine feature mining for video semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Patern Recognition (CVPR)*, 2022.

[43] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

[44] Tsung-Ying Sun, Shang-Jeng Tsai, and Vincent Chan. Hsi color model based lane-marking detection. In *IEEE Intelligent Transportation Systems Conference*, pages 1168–1172. IEEE, 2006.

[45] Lucas Tabelini, Rodrigo Berriel, Thiago M Paixao, Claudine Badue, Alberto F De Souza, and Thiago Oliveira-Santos. Keep your eyes on the lane: Real-time attention-guided lane detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 294–302, 2021.

[46] Lucas Tabelini, Rodrigo Berriel, Thiago M Paixao, Claudine Badue, Alberto F De Souza, and Thiago Oliveira-Santos. Polylanenet: Lane estimation via deep polynomial regression. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6150–6156. IEEE, 2021.

[47] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[48] Luo-Wei Tsai, Jun-Wei Hsieh, Chi-Hung Chuang, and Kuo-Chin Fan. Lane detection using directional random walks. In *IEEE Intelligent Vehicles Symposium*, pages 303–306. IEEE, 2008.

[49] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[51] Jun Wang, Tao Mei, Bin Kong, and Hu Wei. An approach of lane detection based on inverse perspective mapping. In *International IEEE Conference on Intelligent Transportation Systems*, pages 35–38. IEEE, 2014.

[52] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[53] Ruochen Yin, Yong Cheng, Huapeng Wu, Yuntao Song, Biao Yu, and Runxin Niu. Fusionlane: Multi-sensor fusion for lane marking semantic segmentation using deep neural networks. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[54] Seungwoo Yoo, Hee Seok Lee, Heesoo Myeong, Sungrack Yun, Hyoungwoo Park, Janghoon Cho, and Duck Hoon

Kim. End-to-end lane marker detection via row-wise classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1006–1007, 2020.

[55] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

[56] Yujun Zhang, Lei Zhu, Wei Feng, Huazhu Fu, Mingqian Wang, Qingxia Li, Cheng Li, and Song Wang. Vil-100: A new dataset and a baseline model for video instance lane detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15681–15690, 2021.

[57] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.