Video Instance Lane Detection via Deep Temporal and Geometry Consistency Constraints

Mingqian Wang*
College of Intelligence and Computing, Tianjin University
wangmingqian@tju.edu.cn

Yujun Zhang*
College of Intelligence and Computing, Tianjin University
yujunzhang@tju.edu.cn

Wei Feng†
College of Intelligence and Computing, Tianjin University
wfeng@ieee.org

Lei Zhu
The Hong Kong University of Science and Technology (Guangzhou) & The Hong Kong University of Science and Technology
leizhu@ust.hk

Song Wang
University of South Carolina
Columbia, USA
songwang@cec.sc.edu

ABSTRACT

Video instance lane detection is one of the most important tasks in autonomous driving. Due to the very sparse region and weak context in lane annotations, accurately detecting instance-level lanes in real-world traffic scenarios is challenging, especially for scenes with occlusion, bad weather conditions, dim or dazzling lights. Current methods mainly address this problem by integrating features of adjacent video frames to simply encourage temporal constancy for image-level lane detectors. However, most of them ignore lane shape constraint of adjacent frames and geometry consistency of individual lanes, thereby harming the performance of video instance lane detection. In this paper, we propose TGC-Net via temporal and geometry consistency constraints for reliable video instance lane detection. Specifically, we devise a temporal recurrent feature-shift aggregation module (T-RESA) to learn spatio-temporal lane features along horizontal, vertical, and temporal directions of the feature tensor. We further impose temporal consistency constraint by encouraging spatial distribution consistency among the lane features of adjacent frames. Besides, we devise two effective geometry constraints to ensure the integrity and continuity of lane predictions by leveraging pairwise point affinity loss and vanishing point guided geometric context, respectively. Extensive experiments on public benchmark dataset show that our TGC-Net quantitatively and qualitatively outperforms state-of-the-art video instance lane detectors and video object segmentation competitors. Our code and our results have been released at https://github.com/wmq12345/TGC-Net.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Computer vision**; *Scene understanding*.

KEYWORDS

Video instance lane detection, temporal consistency, geometry consistency, lane integrity, lane continuity, feature shift & aggregation.

ACM Reference Format:
Mingqian Wang, Yujun Zhang, Wei Feng, Lei Zhu, and Song Wang. 2022. Video Instance Lane Detection via Deep Temporal and Geometry Consistency Constraints. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3547914

1 INTRODUCTION

Video instance lane detection (ViLD) is the process of automatically perceiving the shape, position and instance label of multiple lanes to ensure cars to follow the lanes precisely. ViLD facilitates many advanced driving assistance applications, such as trajectory planning and lane departure, and becomes one of the most fundamental and challenging task for autonomous driving. However, robust and accurate lane detection under diverse real-world traffic scenarios is still challenging, especially in harsh scenes with severe occlusion, bad weather conditions, dim or dazzle lights.

Although many early studies focus on image-level lane detection [9, 16, 19, 21, 38], considering car cameras can collect streaming videos, ViLD much better satisfies the real-world requirements of autonomous driving oriented lane detection. Recently, Zhang et al. [37] collected a public video instance lane detection benchmark dataset, VIL-100, and present a multi-level memory aggregation network (MMA-Net) for ViLD. However, state-of-the-art ViLD methods, represented by MMA-Net, mainly concentrate on extending image-level lane detectors to handle videos by integrating temporal consistency of adjacent video frames. Almost all of them neglect lane shape constraint of adjacent frames and geometry consistency of individual lanes, thus degrading the accuracy and robustness of video instance lane detection.

*Equal contribution.
†Corresponding authors.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547914

Note that, lanes usually have specific structure with long and continuous shapes and are always parallel to each other, thus multiple lanes always converge to the same vanishing point in each video frame. The above observations have been explored in image-level lane detection. For individual lane structure, slice-by-slice convolutions are used within lane feature tensors to pass structural message from horizontal and vertical directions respectively [19, 38], to maintain the long and continuous lane structure. First- and second-order structural constraints are used to encourage adjacent lane points consistency and straight line prior, respectively [21]. For multiple lanes, vanishing point detection is found to be a beneficial auxiliary task for image-level lane detection [9]. However, both single lane structural prior and vanishing point guided multi-lanes global constraint have never been explored in ViLD, wherein both temporal consistency and geometry constraint are equally important. For difficult scenarios, e.g., lanes are heavily occluded or damaged, low quality features extracted from single frames cannot support reliable ViLD if without jointly considering temporal consistency and geometry constraint.

To address this problem, in this paper, we propose a new model, namely TGC-Net, via temporal and geometry consistency constraints for reliable ViLD. Specifically, we devise a temporal recurrent feature-shift aggregation module (T-RESA) to learn spatio-temporal lane features along horizontal, vertical, and temporal directions of the feature tensor. As an effective generalization to RESA [38], we slice the feature tensor in vertical and horizontal directions for each frame and apply bi-directional spatial aggregation of lane shape features with different strides. Each spatial aggregation corresponds to a unidirectional temporal aggregation to pass lane shape features to corresponding slice of current frame to its next frames. We further impose temporal consistency constraint by encouraging the spatial distribution consistency among the T-RESA generated lane shape features of adjacent frames. At the same time, we propose two effective geometry constraints to ensure the integrity and continuity of video lane predictions by leveraging pairwise lane point affinity loss and vanishing point guided geometric context, respectively. Both temporal consistency and geometry consistency constraints are integrated as training loss functions of the proposed TGC-Net, to guarantee the accuracy and robustness of ViLD.

The major contributions of this paper are three-fold:

- We propose a novel and effective architecture, TGC-Net, for video instance lane detection (ViLD) by jointly considering temporal and geometry consistency constraints. Compared with lane shape information propagated within single frame and simply encouraging temporal feature constancy, our model is more robust and accurate for harsh real-world traffic scenarios.
- We introduce two simple yet effective geometry consistency constraints to guarantee the lanes integrity and continuity, which can benefit the study for both image-level and video-level lane detection.
- The proposed method achieves the state-of-the-art performance compared with recent ViLD detectors and video object segmentation competitors on the public benchmark dataset VIL-100.

2 RELATED WORK

2.1 Lane Detection

Conventional lane detection exploits hand-crafted low level features or specialized features [2, 11, 25, 33], and usually suffers from poor robustness. Recently, lane detection methods based on deep learning achieve significant success and they can be divided into image-based methods [4, 5, 9, 13, 16, 19, 21, 24, 26, 27, 38] and video-based methods [37]. For image-based methods, LaneNet[16] combines the binary lane segmentation and pixel embeddings, and then uses clustering to generate instance lanes. VPGNet [9] uses a multi-task lane detection network guided by vanishing points. UFSA [21] divides the image into grids and then scans grids for lane locations. SAD [5] employs attention distillation between different CNN layers to capture richer scene contexts. Considered the structural characteristics of lanes, IntRA-KD [4] employs knowledge distillation to transfer structural knowledge from teacher to student models. SCNN [19] propagates sliced feature map between adjacent rows and columns along vertical and horizontal directions. Inspried by SCNN, RESA [38] further passes information in a parallel way to reduce time cost. PolyLaneNet [27] regresses each lane instance using a deep polynomial. LSTR [13] describes lanes with explicit mathematical formulas derived from the road structure and camera poses. In [26] and [24] anchors are predefined, which are then used to regress for the relative coordinates of each lane. Different from the above methods that detect lanes from individual images, in [37] the first lane instance detection dataset was built and a video-based lane detection baseline method MMA-Net was developed by utilizing the features of historical frames to help the segmentation on the current frame. Benefit from integrating features of adjacent video frames, MMA-Net outperforms image-based methods by a large margin. However, MMA-Net totally ignores lane distribution consistency prior among adjacent video frames and geometry consistency prior of each lane instance. Further inspired by previous structure-aware image-based lane detection method [38], in this paper we develop a video instance lane detection network TGC-Net to fully explore both temporal and spatial information.

2.2 Video Object Segmentation

The classical video object segmentation (VOS) methods are also applicable for ViLD. The former segments for general objects, while the latter focuses on the specific structure of lanes, which has long and thin structural property. Early video object segmentation methods rely on the temporal label propagation [1, 10, 17] based on heuristic cues. These traditional methods cannot effectively deal with the challenging situations of severe occlusion, large displacement and so on. Recent deep-learning based methods have shown great advantages and they can be divided into zero-shot methods [6, 7, 23, 32, 34, 39] and one-shot methods [8, 12, 18, 22, 35, 36]. They differ in that the former does not require the true instance segmentation of the first frame while the latter does. For zero-shot methods, object saliency was exploited in [6, 23] and motion and appearance were integrated in [7, 39]. In [32], both spatial and temporal consistency are considered within frames. VisTR [34] verifies the effectiveness of Transformer[31] in video object segmentation. For the one-shot VOS, GAM [8] learns the appearance of targets and background in a single forward pass to avoid online

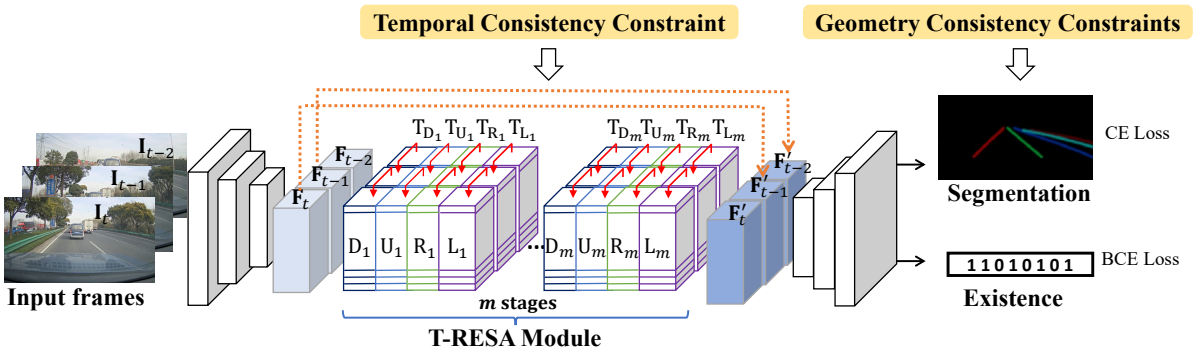


Figure 1: Schematic illustration of the proposed TGC-Net. Overall architecture is composed by Encoder, T-RESA module and Decoder. 'D', 'U', 'R', and 'L' denote downward, upward, rightward, and leftward information passing respectively per stage in T-RESA module for each video frame. These four direction information passing within a frame combined with synchronous passing across frames complete a spatio-temporal information propagation. We repeat m stages in T-RESA module. m is set to be 5 in our method.

fine-tuning. To cope with object occlusion and drifting, STM [18] leverages a memory network and performs pixel-level matching to find objects that are similar to the target objects from past frames. Based on STM [18], KMN [22] applies a Gaussian kernel to reduce the mismatched pixels, RMNet [35] uses a local-to-local matching based on the optical flow estimated from the previous frame. The accuracy of segmentation can be further improved by enhancing the quality of boundary segmentation [12]. TVOS [36] follows a label propagation approach by combining the annotation of the first frame and historical frames. The above methods verify that temporal continuity provides a strong clue for video tasks. Especially in video instance lane detection, car cameras always take streaming videos on the way and it is highly desirable to leverage temporal consistency in dynamic videos to resolve the in-frame ambiguities that are common in image-based lane detection. Moreover, continuous video instance lane detection is indispensable for advanced driving assistance system (ADAS) [14, 15, 28], such as lane change, trajectory planning, and autonomous navigation. Motivated by this, we propose a new TGC-Net equipped with temporal and structural consistency constraints among adjacent video frames in this paper.

3 OUR APPROACH

3.1 Overview

Figure 1 shows the schematic illustration of our video instance lane detection network (TGC-Net), which explores more temporal/structural consistencies between adjacent video frames and geometry consistency prior of each individual lane. For current (target) frame $\mathbf{I}_t$, our TGC-Net starts with an ordered video sequence $\mathbf{I}_t, \mathbf{I}_{t-1}, \cdots, \mathbf{I}_{t-T+1}$ drawn from the input lane video stream, i.e., the current frame and $T-1$ historical frames just before that. To trade-off between computation efficiency and accuracy, T is set to be 3 in our experiments. These ordered frames are fed into CNN Encoder (ResNet) to obtain initial feature maps $\mathbf{F}_s \in \mathbb{R}^{C\times H\times W}$, $s = t, t-1, \cdots, t-T+1$, where C, H, and W denote the channel, height, and width respectively. We then develop a Temporal REcurrent Feature-Shift Aggregator (T-RESA) to gather spatial temporal information synchronously by shifting sliced feature map horizontally and vertically, and passing structure information along frames. This way, we get corresponding spatial-temporal aggregated structural features $\mathbf{F}'_s \in \mathbb{R}^{C\times H\times W}$, thus we can effectively propagate slice-based regional temporal consistency information along ordered video sequence. Besides, to propagate image-based global temporal consistency information along frames, such as similar lane distribution among adjacent frames, we impose a temporal consistency constraint to ensure that $\mathbf{F}_s$ get richer context message from enhanced feature map $\mathbf{F}'_{s-1}$, $s = t, t-1, \cdots, t-T+2$ from last frame. Then we adopt a bilateral up-sampling decoder [38] on the final spatial-temporal aggregated feature map $\mathbf{F}'_t$ to predict video lane instance results for the target frame $\mathbf{I}_t$. Finally, we introduce geometry consistency constraints to enhance the integrity and continuity in lane instance segmentation.

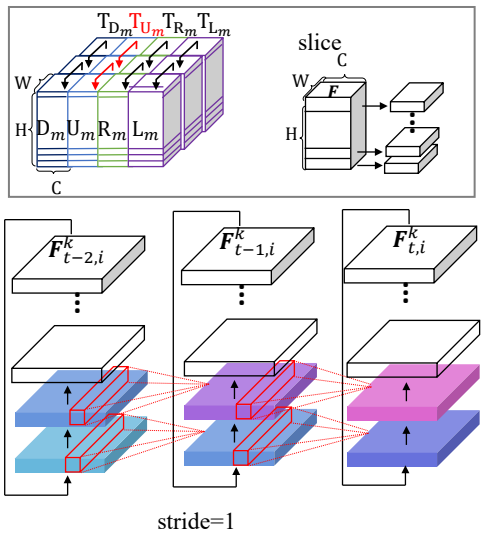


Figure 2: Schematic illustration of upward Temporal REcurrent Feature-Shift Aggregator (T-RESA) when stide is 1 for m-th stage. The propagation for other three directions are similar.

3.2 T-RESA Module

Temporal REcurrent feature-Shift Aggregator (T-RESA). REcurrent Feature-Shift Aggregator (RESA) [38] aggregates spatial information of a single 2D image in an iterative manner. In each stage, the input feature map of RESA is sliced along the H and W dimensions. At the H dimension, the sliced feature map shifts recurrently in two directions (i.e. downward and upward), and passes information vertically. Similarly, the sliced feature map along the W dimension propagates information in two horizontal directions (i.e., rightward and leftward). To adapt it for learning video features, we propose a Temporal REcurrent Feature-Shift Aggregator (T-RESA) to effectively integrate the temporal information encoded in video frames into RESA for addressing video instance lane detection. As shown in Figure 1, apart from the horizontal (i.e., W) and vertical (i.e., H) directions in the original RESA, our T-RESA shifts the sliced input feature map along the temporal direction of adjacent video frames.

Figure 2 shows the schematic illustration of T-RESA along the upward direction at one stage. The upward propagation takes three feature maps $\mathbf{F}_{t-2}$, $\mathbf{F}_{t-1}$, and $\mathbf{F}_t$ from input T (T=3) adjacent video frames. The size of each feature map is $C\times H\times W$, where C, H, and W denote the number of channels, rows, and columns respectively. We slice each feature map into H slices and thus the size of each slice is $C\times 1\times W$ along the H dimension. Unlike [19], we pass information among slices in a parallel way as in [38], thereby resulting in a time-saving information propagation by several iterations with different stride values. Specifically, in iteration k, the sliced feature maps from $\mathbf{F}_{t-2}$ only vertically propagate information among slices of itself, while the sliced feature maps from $\mathbf{F}_{t-1}$ and $\mathbf{F}_t$ not only receive the information from their own frame, but also receive information from the last frames, e.g., when stride is 1, the sliced feature map $\mathbf{F}^k_{t,i}$, where t, i, and k indicate the indices of frame, row, and iteration respectively, receives the information from both $\mathbf{F}^k_{t,i-1}$ and $\mathbf{F}^k_{t-1,i}$. We formulate the information propagation in a given direction as

$$\begin{cases} \mathbf{F}^{k'}_{t,i} = \mathbf{F}^k_{t,i} + f^k_i(\mathbf{F}^k_{t,(i-s_k) \bmod N}) & t = 0 \\ \mathbf{F}^{k'}_{t,i} = \mathbf{F}^k_{t,i} + f^k_i(\mathbf{F}^k_{t,(i-s_k) \bmod N}) + g^k_i(\mathbf{F}^k_{t-1,(i-s_k) \bmod N}) & t \geq 1 \\ s_k = \frac{N}{2^K-k},\ k = 1, 2, 3, \cdots, K. \end{cases} \tag{1}$$

Here k denotes the index of iteration as in [38], and the iteration number $K = \lfloor \log_2 N \rfloor$ with N being the number of slices, i.e., N equals to H for vertical propagation, and W for horizontal propagation. $i \in \{0, 1, 2, \cdots, N-1\}$ and t are the the indices of slices and frames, respectively. s_k is the stride for iteration k. $f^k_i(\cdot)$ consists of 1-d convolution kernel and nonlinear activation function ReLU. 1-d convolution kernal size is $C_{\text{in}} \times C_{\text{out}} \times w$, where C_{in}, C_{out}, and w denote the number of input channels, the number of output channels, and the kernel width respectively. Both C_{in} and C_{out} are the same as C. The parameters of 1-d convolution kernels are shared in different frames. $g^k_i(\cdot)$ is similar to $f^k_i(\cdot)$, and the parameters of 1-d convolution kernels are shared between adjacent frames. $F^{k'}_{t,i}$ denote the updated sliced feature map.

Temporal Consistency constraint. Although T-RESA can learn temporal information, it only propagates temporal consistency information in slice-based local regions. To leverage image-based global lane shape consistency priors between adjacent video frames, we introduce a temporal consistency constraint between initial feature $\mathbf{F}_s$ and enhanced feature $\mathbf{F}^{'}_{s-1}$ through T-RESA module. Specifically, we define an activation mapping function $\mathcal{G}(\mathbf{X}) = \frac{1}{C}\sum_{c=1}^{C}|\mathbf{X}_c|$, where $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ denotes the input feature map, $\mathbf{X}_c \in \mathbb{R}^{H\times W}$ is c-th channel feature map of the input. We impose the activation mapping function $\mathcal{G}(\cdot)$ on both $\mathbf{F}_s$ and $\mathbf{F}^{'}_{s-1}$, then get the corresponding attention maps $\mathbf{Z}_s$ and $\mathbf{Z}^{'}_{s-1}$. To explore richer context information by global temporal consistency, we encourage $\mathbf{Z}_s$ to mimic $\mathbf{Z}^{'}_{s-1}$. The temporal consistency constraint $\mathcal{L}_{\text{T}}$ is defined as:

$$\mathcal{L}_{\text{T}} = \sum_{s=t-T+2}^{t} \Phi_{\cos}(Z_s, Z^{'}_{s-1}), \tag{2}$$

where $\Phi_{\cos}(\cdot)$ denotes the cosine embedding loss.

3.3 Geometry Consistency Constraints

Lane integrity constraint. To maintain the integrity of each instance lane, inspired by [29], we define a pairwise point affinity loss according to ground-truth labels. As shown in Figure 3 (b), if point x_i and its adjacent eight neighbors $\{x_j\}$ have the same categorical label in ground truth, the point x_i is called an inner point, and we encourage the prediction results of x_i and its neighbor x_j to be consistent, which is named structural integrity constraint in our work. Otherwise, the point x_i is called an edge point, and we push apart their prediction results.

Specifically, as shown in Figure 3 (c), for a lane instance category c, the ground-truth probability of the point x_i which belongs to the instance c satisfies Bernoulli distribution $P(x_i) = [p_i, 1-p_i]$, where $p_i = \{1, 0\}$, and the prediction probability $Q(x_i) = [q_i, 1-q_i]$, where $q_i \in (0, 1)$. Similarly, the ground-truth and prediction probability of neighbor point x_j are denoted as $P(x_j)$ and $Q(x_j)$. According to ground-truth, we define the indices of edge point as $\mathbf{I}_{\text{edge}} = \{i | P(x_i) \neq P(x_j)\}$, and the indices of inner point as $\mathbf{I}_{\text{inner}} = \overline{\mathbf{I}_{\text{edge}}}$. The pairwise point affinity loss is formulated as:

$$\mathcal{L}_{\text{GI}} = \begin{cases} D_{\text{KL}}(Q(x_j)||Q(x_i)) & i \in \mathbf{I}_{\text{inner}} \\ \max\{0, m - D_{\text{KL}}((Q(x_j)||Q(x_i))\} & i \in \mathbf{I}_{\text{edge}}, \end{cases} \tag{3}$$

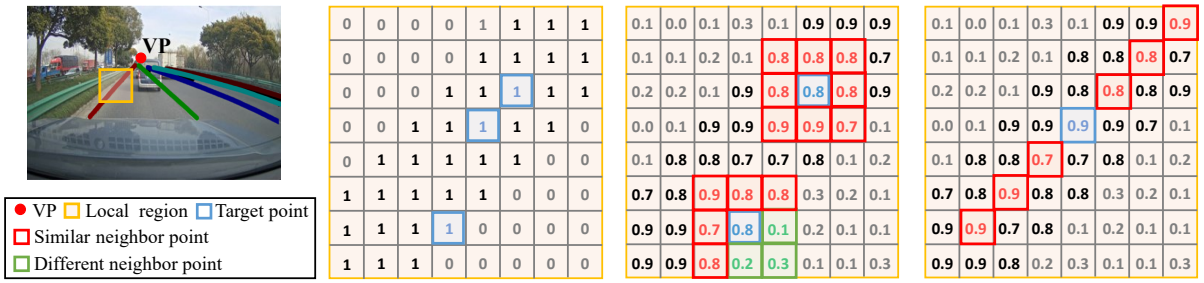
where $D_{\text{KL}}(Q(x_j)||Q(x_i)) = q_j \log\frac{q_j}{q_i} + \overline{q_j}\log\frac{\overline{q_j}}{\overline{q_i}}$ is the Kullback-Leibler divergence between $Q(x_j)$ and $Q(x_i)$. As shown by considered point (in blue box) in Figure 3 (c), we shall not pull it apart from all eight adjacent point in segmentation – we only need to pull it apart from background. Considering this, we further update Eq. (3) to

$$\mathcal{L}_{\text{GI}} = \begin{cases} D_{\text{KL}}(Q(x_j)||Q(x_i)) & i \in \mathbf{I}_{\text{inner}} \\ (p_i \oplus p_j) \odot \max\{0, m - D_{\text{KL}}((Q(x_j)||Q(x_i))\} & i \in \mathbf{I}_{\text{edge}}, \end{cases} \tag{4}$$

where $\oplus$ denotes exclusive OR operator, $p_i \oplus p_j$ ensures the point in edge region is pulled apart from neighbors in a certain direction with different labels, and $\odot$ means element-wise multiplication.

Lane continuity constraint. In most cases, parallel lanes of video frames converge to a vanishing point (VP); see the red point at Figure3 (a). In this paper, we use VP to provide global geometric context of a scene and ensure the continuity of detected lanes.

For lane instance category c, we encourage the points located along the center line of the lane and vanishing point are on the same line. Here, We employ a simple and efficient method to calculate



(a) ground truth (b) ground truth of local region (c) structural integrity constraint (d) structural continuity constraint

Figure 3: Schematic illustration of geometry consistency constraints. For clarity, we enlarge local region in (a) of a lane instance and show the detailed values in (b-d). (a) Ground truth, the different colors denote instance-level lane annotations and the red point is the vanishing point (VP). (b) Ground truth of local region corresponding to the boxed area in (a). 1 and 0 denote the point belongs to lane instance and background respectively. (c) and (d) are schematic illustration of structural integrity and continuity constraints respectively. Blue box indicates the considered point x_i, and red/green box indicates its similar/different neighbor.

the points along the center line – we sample the point x_i with foreground probability $q_i \in (0, 1)$ greater than 0.5 and remove the outliers, and denote the indices of these points as $\mathbf{I}_{\text{center}}$. Then we calculate slope coefficient of the points along the center line guided by the vanishing point. As shown in Figure 3 (d), we sample six neighbors along and against the slope coefficient direction and define the loss of the structural continuity constraint as

$$\mathcal{L}_{\text{GC}} = D_{\text{KL}}(Q(x_j)||Q(x_i)) \quad i \in \mathbf{I}_{\text{center}}, \tag{5}$$

where $D_{\text{KL}}(Q(x_j)||Q(x_i)) = q_j \log\frac{q_j}{q_i} + \overline{q_j}\log\frac{\overline{q_j}}{\overline{q_i}}$, and it is similar to the one in Eq. (3).

3.4 Implementation Details

Loss function. The total loss $\mathcal{L}_{\text{total}}$ of our TGC-Net consists of a segmentation loss $\mathcal{L}_{\text{Seg}}$, existence loss $\mathcal{L}_{\text{E}}$, and our temporal consistency loss $\mathcal{L}_{\text{T}}$ (see Eq. (2)), and our geometry consistency loss $\mathcal{L}_{\text{G}}$. Moreover, $\mathcal{L}_{\text{G}}$ contains a structural integrity loss $\mathcal{L}_{\text{GI}}$ (see Eq. (4)) and a structural continuity loss $\mathcal{L}_{\text{GC}}$(see Eq. (5)). Hence, the definition of $\mathcal{L}_{\text{total}}$ is given by:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Seg}} + \alpha_1\mathcal{L}_{\text{E}} + \alpha_2\mathcal{L}_{\text{T}} + \alpha_3\mathcal{L}_{\text{G}}\,, \tag{6}$$

where

$$\mathcal{L}_{\text{Seg}} = \Phi_{\text{CE}}(P^t_{\text{seg}}, G^t_{\text{seg}}) \tag{7}$$

$$\mathcal{L}_{\text{E}} = \Phi_{\text{BCE}}(P^t_{\text{exist}}, G^t_{\text{exist}}) \tag{8}$$

$$\mathcal{L}_{\text{G}} = \beta_1\mathcal{L}_{\text{GI}} + \beta_2\mathcal{L}_{\text{GC}}\,. \tag{9}$$

Here Φ_{CE} and Φ_{BCE} denote the cross entropy loss function and the binary cross entropy loss function [19, 38], respectively. P^t_{seg} and G^t_{seg} represent the instance lane prediction map and the corresponding ground truth for $t-th$ input (current) frame. P^t_{exist} and G^t_{exist} denote the binary classification for each lane instance and the corresponding ground truth. We empirically set the weights α_1, α_2, and α_2 as: $\alpha_1 = 0.1$, $\alpha_2 = 0.2$, and $\alpha_3 = 0.4$. And the weights β_1 and β_2 are empirically set as $\beta_1 = 1.0$ and $\beta_2 = 1.0$ in our experiments.

Network Training. In the preprocessing stage, considering that VIL-100 has no annotated vanishing point, we take the intersection point of multiple lane curves (obtained by a straight line fitting) in each video frame as the ground truth of the vanishing point. We also utilize a random crop operation on each current video frame for data augmentation. our TGC-Net is based on an encoder-decoder architecture, where the feature extraction backbone at the encoder is ResNet-50 [3], and the decoder is the bilateral up-sampling decoder proposed in [38]. Note that the geometry consistency constraints requires a suitable video lane detection result. Hence, we empirically utilize two stages to train our TGC-Net. In the first training stage, we remove the geometry consistency constraints $\mathcal{L}_{\text{GC}}$ from the total loss (see Eq. 3.4) to train our TGC-Net. In the second training stage, we fine-tune the train model in the first stage by containing all components of the total loss of our TGC-Net.

Training parameters. We implement our TGC-Net using Pytorch and train our network on a NVIDIA GTX 3090Ti. In the first training stage, we initialize the feature extraction backbone by using a pre-trained ResNet-50 [3], and employ a SGD optimizer with a learning rate of 1×10^{-2}, a momentum of 9×10^{-1}, a weight decay of 1×10^{-4}, and mini-batch size of 4. In the second stage, $\mathcal{L}_{\text{GC}}$ is employed to fine-tune the whole network with a learning rate as 1×10^{-4}, a momentum of 9×10^{-1}, a weight decay of 1×10^{-4} and mini-batch size of 2. The first training stage requires 40 epochs, while the second training stage requires 20 epochs.

4 EXPERIMENTAL RESULTS

Dataset and Evaluation Metrics We evaluate our network and compared state-of-the-art methods on a public video instance lane detection dataset (i.e., VIL-100 [37]). VIL-100 consists of 100 videos, 10, 000 frames, and each frame of all videos are with instance-level annotations on lane regions. The dataset contains 10 typical scenarios: normal, crowded, curved road, damaged road, shadows, road markings, dazzle light, haze, night, crossroad. The latter nine scenarios are common and challenging. The dataset also provides two type of annotations: point annotations along the center line and region annotations of each lane.

Following the existing method [37], we first employ six widely-used image-based metrics, including three region-based metrics and three line-based metrics. Three region-based metrics [19] are mIoU, F1(IoU>0.5) (denoted as $\text{F1}^{0.5}$), and F1(IoU>0.8) (denoted as

Table 1: Quantitative comparisons of our network and state-of-the-art methods in terms of image-based metrics.

Methods	Year	Region			Line		
		mIoU↑	$F1^{0.5}$ ↑	$F1^{0.8}$ ↑	Accuracy ↑	FP ↓	FN ↓
LaneNet [16]	2018	0.633	0.721	0.222	0.858	0.122	0.207
SCNN [19]	2018	0.517	0.491	0.134	0.907	0.128	0.110
ENet-SAD [5]	2019	0.616	0.755	0.205	0.886	0.170	0.152
UFSA [21]	2020	0.465	0.310	0.068	0.852	0.115	0.215
LSTR [13]	2021	0.573	0.703	0.131	0.884	0.163	0.148
LaneATT [26]	2021	0.452	0.337	0.061	0.712	0.093	0.357
RESA [38]	2021	0.702	0.874	0.345	0.936	0.078	0.068
MMA-Net [37]	2021	0.705	0.839	0.458	0.910	0.111	0.105
GAM [8]	2019	0.602	0.703	0.316	0.855	0.241	0.212
RVOS [32]	2019	0.294	0.519	0.182	0.909	0.610	0.119
STM [18]	2019	0.597	0.756	0.327	0.902	0.228	0.129
AFB-URR [12]	2020	0.515	0.600	0.127	0.846	0.255	0.222
TVOS [36]	2020	0.157	0.240	0.037	0.461	0.582	0.621
VisTR [34]	2021	0.683	0.841	0.303	0.917	0.097	0.101
RMNet [35]	2021	0.667	0.807	0.402	0.897	0.163	0.132
TGC-Net (ours)	2022	**0.738**	**0.892**	**0.469**	**0.941**	**0.064**	**0.057**

Table 2: Quantitative comparisons of our network and state-of-the-art methods in terms of video-based metrics.

Methods	$\mathcal{M}_{\mathcal{J}}\uparrow$	$\mathcal{O}_{\mathcal{J}}\uparrow$	$\mathcal{M}_{\mathcal{F}}\uparrow$	$\mathcal{O}_{\mathcal{F}}\uparrow$	$\mathcal{M}_{\mathcal{J}\&\mathcal{F}}\uparrow$
GAM [8]	0.414	0.203	0.721	0.781	0.568
RVOS [32]	0.251	0.251	0.251	0.251	0.251
STM [18]	0.656	0.626	0.743	0.763	0.656
AFB-URR [12]	0.308	0.251	0.415	0.435	0.362
TVOS [36]	0.255	0.251	0.257	0.256	0.255
VisTR [34]	0.696	0.742	0.849	0.893	0.773
RMNet [35]	0.499	0.592	0.678	0.694	0.588
MMA-Net [37]	0.679	0.735	0.848	0.873	0.764
TGC-Net (Ours)	**0.727**	**0.794**	**0.891**	**0.906**	**0.809**

$F1^{0.8}$), while three line-based metrics are Accuracy, FP, and FN [30]. In general, a better video instance lane detection method shall have larger mIoU, $F1^{0.5}$, $F1^{0.8}$, and Accuracy scores, as well as smaller FP and FN scores. Moreover, we also employ five video-based metrics to evaluate the temporal stability of the video segmentation results [20], i.e., $\mathcal{M}_{\mathcal{J}}$, $\mathcal{O}_{\mathcal{J}}$, $\mathcal{M}_{\mathcal{F}}$, $\mathcal{O}_{\mathcal{F}}$, and $\mathcal{M}_{\mathcal{J}\&\mathcal{F}}$, a better video instance segmentation method should have larger scores for all five video-based metrics.

Comparative methods. To demonstrate the superiority of our video instance lane detection method, we compare it against 15 state-of-the-art methods, including LaneNet [16], SCNN [19], ENet-SAD [5], UFSA [21], LSTR [13], LaneATT [26], RESA [38], MMA-Net [37], GAM [8], RVOS [32], STM [18], AFB-URR [12], TVOS [36], VisTR [34], and RMNet [35]. Among them, LaneNet, SCNN, ENet-SAD, UFSA, LSTR, LaneATT, and RESA are image-level lane detection methods, while MMA-Net is video-level lane detection method. Moreover, GAM, RVOS, STM, AFB-URR, TVOS, VisTR, and RMNet are instance-level video object segmentation methods.

For LaneATT, RESA, VisTR and RMNet, we use their public implementations, and re-train these methods on video instance-level lane detection dataset VIL-100 to obtain their best results for a fair comparison. For one-shot VOS methods (such as [35]), we utilize the ground truth of the first frame in each video clip to train. And the results of other methods are directly from MMA-Net [37].

4.1 Comparisons with State-of-the-art Methods

Quantitative comparisons. Table 1 reports six image-level quantitative results of our network and all compared methods. We can observe that lane detection methods tend to have a better performance on both line and region based metrics compare than general video object segmentation (VOS) methods. The reason is that the lanes are slender and continuous, and lane detection methods can utilize structural information (such as fitting curves[27] or predicting vanishing points[9]) to detect lanes, while the video lane detection method MMA-Net and all VOS methods pay more attention to temporal continuity and totally ignore the shape and direction of lanes. Specifically, among all compared methods, MMA-Net has the best mIoU score of 0.705 and the best $F1^{0.8}$ of 0.458, while RESA has the best $F1^{0.5}$ of 0.874, the best Accuracy score of 0.936, the best FP score of 0.078, the best FN score of 0.068. Compared to all best metric scores of compared methods, our network has a mIoU improvement of 4.68%, a $F1^{0.5}$ improvement of 2.06%, a $F1^{0.8}$ improvement of 2.40%, a Accuracy improvement of 0.53%, a FP improvement of 17.95%, and a FN improvement of 16.18%.

To further explain the effectiveness of our method, we list video-based metric scores of our network and other video object segmentation methods in Table 2. From these video-based metric results, we can find that VisTR has the largest score of all video-based metrics among compared methods, they are $\mathcal{M}_{\mathcal{J}}$ of 0.696, $\mathcal{O}_{\mathcal{J}}$ of 0.742, $\mathcal{M}_{\mathcal{F}}$ of 0.849, $\mathcal{O}_{\mathcal{F}}$ of 0.893 and $\mathcal{M}_{\mathcal{J}\&\mathcal{F}}$ of 0.773. More importantly, our method outperforms VisTR on all video-based metrics. Specifically, our method improves $\mathcal{M}_{\mathcal{J}}$ from 0.696 to 0.727; $\mathcal{O}_{\mathcal{J}}$ from 0.742 to 0.794; $\mathcal{M}_{\mathcal{F}}$ from 0.849 to 0.891; $\mathcal{O}_{\mathcal{F}}$ from 0.893 to 0.906; and $\mathcal{M}_{\mathcal{J}\&\mathcal{F}}$ from 0.773 to 0.809. It indicates that our network can more accurately detect instance-level lanes than compared methods.

Visual comparisons. Figure 4 visually compares video instance lane detection results produced by our TGC-Net and state-of-the-art methods. For image-based lane detection methods, as shown in 5-th to 8-th rows, they often neglect lane regions when these lanes are severe occluded and damaged in the road. For video object segmentation methods at 9-th to 11-th rows, the lane instances are

Table 3: Ablation study of our proposed method in terms of image-level and video-level metrics.

Network	Component				Image-based metrics						video-based metrics				
	T-RESA	$\mathcal{L}_{\mathrm{T}}$	$\mathcal{L}_{\mathrm{GI}}$	$\mathcal{L}_{\mathrm{GC}}$	mIoU↑	$\mathrm{F1}^{0.5}$ ↑	$\mathrm{F1}^{0.8}$ ↑	Accuracy ↑	FN ↓	FP ↓	$\mathcal{M}_{\mathcal{J}}$ ↑	$\mathcal{O}_{\mathcal{J}}$ ↑	$\mathcal{M}_{\mathcal{F}}$ ↑	$\mathcal{O}_{\mathcal{F}}$ ↑	$\mathcal{M}_{\mathcal{J}\&\mathcal{F}}$ ↑
"basic-TR"	✓				0.716	0.881	0.393	0.940	0.075	0.058	0.722	0.775	0.886	0.901	0.804
"basic-TR+TC"	✓	✓			0.720	0.889	0.400	0.941	0.073	0.058	0.722	0.781	0.886	0.903	0.804
"basic-TR+TC+GI"	✓	✓	✓		0.732	0.891	0.427	0.937	0.065	0.063	0.717	0.768	0.879	0.898	0.798
"basic-TR+TC+GI+GC" (ours)	✓	✓	✓	✓	0.738	0.892	0.469	0.941	0.064	0.057	0.727	0.794	0.891	0.906	0.809

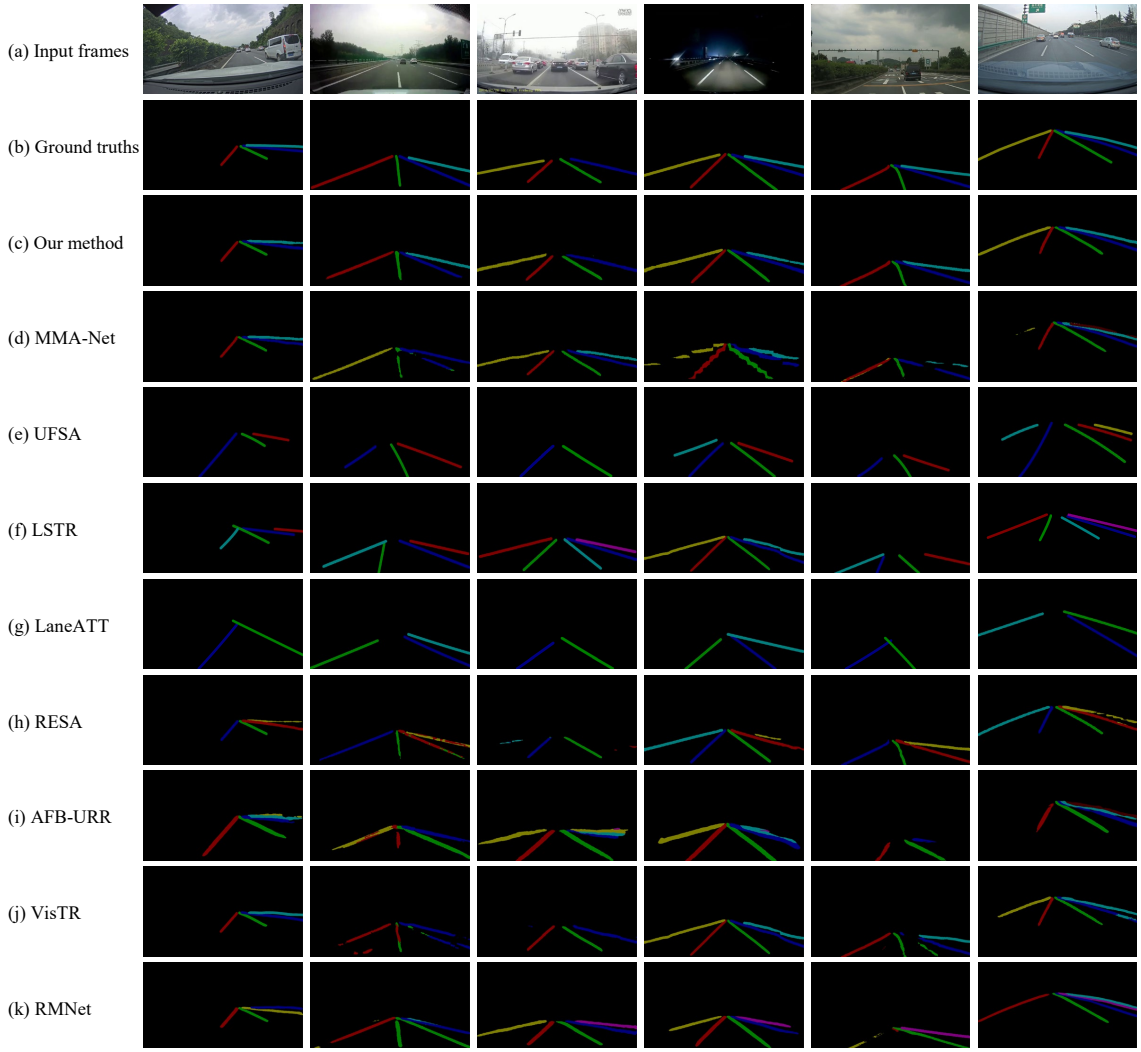


Figure 4: Visual comparisons of video instance lane detection maps produced by our network (3rd row) and state-of-the-art methods (4-th to 11-th rows). The ground truths is shown at the 2nd row. Apparently, our network can more accurately identify instance-level lanes than all compared state-of-the-art methods.

also mistakenly detected or missed, and their lane detection results are worse than lane detection methods. The detected lanes of MMA-Net are not slender and smooth, and its segmentation results of single lane are not continuous; see the 4-th row of Figure 4. On the contrary, our method at the 3rd row can accurately identify lanes, especially in challenging cases, such as occluded lane regions. Moreover, our lane detection results can well maintain the integrity and consistency of the lanes. In addition, Figure 5 presents additional comparisons results of our TGC-Net, MMA-Net, and RESA in several challenging cases, e.g., occlusion, bad weather, dim and dazzling lights. We can clearly see that detected lane instances of MMA-Net and RESA are not slender and continuous, while our method can better detect all instance-level lanes in different challenging video frames.

4.2 Ablation Study

Basic network design. Here, we construct three baseline networks to evaluate the effectiveness of four major components of our method, and they are "T-RESA", "$\mathcal{L}_{\mathrm{T}}$", "$\mathcal{L}_{\mathrm{GI}}$", and "$\mathcal{L}_{\mathrm{GC}}$". Specifically, "T-RESA" is temporal recurrent feature-shift aggregator to

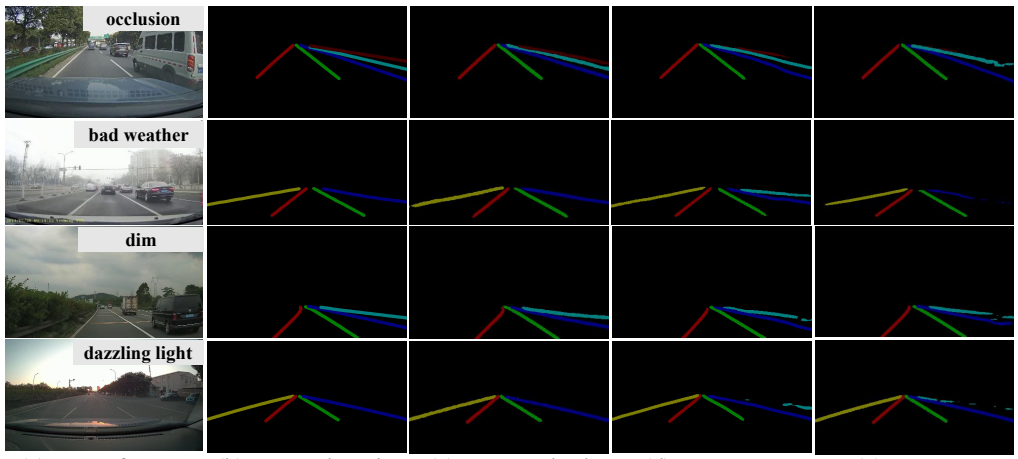


Figure 5: More visual comparisons of our method and state-of-the-art methods(MMA-Net [37] and RESA [38]) in scenes with occlusion, bad weather conditions, dim or dazzling lights.

gather spatial temporal information synchronously. "$\mathcal{L}_{\mathrm{T}}$", "$\mathcal{L}_{\mathrm{GI}}$", and "$\mathcal{L}_{\mathrm{GC}}$" are the temporal consistency constraint, the structural integrity constraint, and the structural continuity constraint respectively. As shown in Table 3, The first baseline network (denoted as "basic-TR") only contains "T-RESA" of our method. It equals to remove the temporal and geometry consistency constraints from our network. The second baseline (denoted as "basic-TR+TC") is to add the temporal consistency constraint "$\mathcal{L}_{\mathrm{T}}$" into the first baseline network. It removes the geometry consistency constraints from our network TGC-Net. The third baseline network denoted as "basic-TR+TC+GI" is to add the structural integrity constraint "$\mathcal{L}_{\mathrm{GI}}$" into the second baseline. Table 3 reports image-based and video-based metric results of our method and all constructed three baseline networks. Please refer to supp. material for visualization comparisons.

Effectiveness of the T-RESA Module. As shown in Table 3, "basic-TR" outperforms "RESA"[38](see Table 1) on all image-level and video-level metrics. It demonstrates that embedding the video temporal information into "RESA" [38] via our "T-RESA" improves the video instance lane detection accuracy.

Effectiveness of the Temporal Consistency Constraint. Furthermore, we can find that "basic-TR+TC" has a larger mIoU score , a larger $F1^{0.8}$ score, a larger $F1^{0.5}$ score, a larger Accuracy score, a smaller FN score, a smaller FP score, a larger $\mathcal{M}_{\mathcal{J}}$ score, a larger $O_{\mathcal{J}}$ score, a larger $\mathcal{M}_{\mathcal{F}}$ score, a larger $O_{\mathcal{F}}$ score, and a larger $\mathcal{M}_{\mathcal{J}\&\mathcal{F}}$ score than "basic-TR". It demonstrates that integrating image-based global priors from adjacent video frames via our temporal consistency constraint can enhance the instance lane segmentation results of the target video frame.

Effectiveness of the Geometry Consistency Constraints. Note that our geometry consistency constraint contains a structural integrity constraint "$\mathcal{L}_{\mathrm{GI}}$" and a structural continuity constraint "$\mathcal{L}_{\mathrm{GC}}$". By observing Table 3, we can find that "basic-TR+TC+GI" has superior performance in region related image-based metrics (i.e.mIoU, $F1^{0.8}$, $F1^{0.5}$) than "basic-TR+TC". However, since "GI" only considers local geometry with eight neighbors but not considers long structure constraints, "basic-TR+TC+GI" tends to produce lower results in some metrics. Therefore, we further introduce lane continuity constraint (denoted as "GC"), it considers the global geometric context of a scene based on vanishing point. By doing so, our method (denoted as "basic-TR+TC+GI+GC") consistently outperforms "basic-TR", "basic-TR+TC" and "basic-TR+TC+GI" on all metrics. It also indicates the structural continuity constraint "$\mathcal{L}_{\mathrm{GC}}$" is capable to further improve the video instance lane detection performance of our method by promoting the continuity of lanes and maintaining the slender structure of lanes.

5 CONCLUSION

In this paper, we have proposed TGC-Net, a new and effective model for reliable video instance lane detection (ViLD), via temporal and geometry consistency constraints. Specifically, we present a genearlized temporal recurrent feature-shift aggregation module (T-RESA) to learn spatio-temporal lane shape features along horizontal, vertical, and temporal directions from the feature tensor. T-RESA propagates temporal consistent information in slice-based local region, therefore we imposes temporal shape consistency constraints on features extracted from neighboring video frames to leverage the image-based global strong lane shape priors among adjacent frames. We further impose temporal consistency constraint by encouraging spatial distribution consistency among the lane features of adjacent frames. Additionally, we devise two effective geometry constraints to ensure the integrity and continuity of lane predictions by leveraging pairwise point affinity loss and vanishing point guided geometric context, respectively. Our method is evaluated on public video instance lane detection benchmark dataset, i.e., VIL-100, and have achieved the superior performance over state-of-the-art competitors. In the future, we want to study the 3D ViLD problem and apply the TGC-Net to other line-shape object detection tasks, e.g., power-line-cruising.

Acknowledgments: The work is supported by the National Natural Science Foundation of China (Grant No. 62072334, 61902275, U1803264).

REFERENCES

[1] Alon Faktor and Michal Irani. 2014. Video Segmentation by Non-Local Consensus Voting. In *BMVC*, Vol. 2. 8.

[2] Chao Fan, Li Long Hou, Shuai Di, and Jing Bo Xu. 2012. Research on the Lane Detection Algorithm Based on Zoning Hough Transformation. In *Advanced Materials Research*, Vol. 490. Trans Tech Publ, 1862–1866.

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

[4] Yuenan Hou, Zheng Ma, Chunxiao Liu, Tak-Wai Hui, and Chen Change Loy. 2020. Inter-region affinity distillation for road marking segmentation. In *CVPR*. 12486–12495.

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

[6] Yuanting Hu, Jiabin Huang, and Alexander G Schwing. 2018. Unsupervised Video Object Segmentation using Motion Saliency-Guided Spatio-Temporal Propagation. In *ECCV*. 786–802.

[7] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. 2017. FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *CVPR*. IEEE, 2117–2126.

[8] Joakim Johnander, Martin Danelljan, Emil Brissman, Fahad Shahbaz Khan, and Michael Felsberg. 2019. A Generative Appearance Model for End-to-end Video Object Segmentation. In *CVPR*. 8953–8962.

[9] Seokju Lee, Junsik Kim, Jae Shin Yoon, Seunghak Shin, Oleksandr Bailo, Namil Kim, Tae-Hee Lee, Hyun Seok Hong, Seung-Hoon Han, and So Kweon. 2017. Vpgnet: Vanishing point guided network for lane and road marking detection and recognition. In *ICCV*. 1947–1955.

[10] Yong Jae Lee, Jaechul Kim, and Kristen Grauman. 2011. Key-Segments for Video Object Segmentation. In *ICCV*. IEEE, 1995–2002.

[11] Zuo-Quan Li, Hui-Min Ma, and Zheng-Yu Liu. 2016. Road Lane Detection With Gabor filters. In *ISAI*. IEEE, 436–440.

[12] Yongqing Liang, Xin Li, Navid Jafari, and Jim Chen. 2020. Video Object Segmentation with Adaptive Feature Bank and Uncertain-Region Refinement. *NeurIPS* 33 (2020), 3430–3441.

[13] Ruijin Liu, Zejian Yuan, Tie Liu, and Zhiliang Xiong. 2021. End-to-end lane shape prediction with transformers. In *WACV*. 3694–3702.

[14] Nicolás Madrid and Petr Hurtik. 2016. Lane Departure Warning for mobile devices based on a fuzzy representation of images. *Fuzzy Sets and Systems* 291 (2016), 144–159.

[15] Joel C McCall and Mohan M Trivedi. 2006. Video-Based Lane Estimation and Tracking for Driver Assistance: Survey, System, and Evaluation. *IEEE Trans. Intell. Transp. Syst.* 7, 1 (2006), 20–37.

[16] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. 2018. Towards End-to-End Lane Detection: an Instance Segmentation Approach. In *IV*. IEEE, 286–291.

[17] Peter Ochs and Thomas Brox. 2011. Object Segmentation in Video: A Hierarchical Variational Approach for Turning Point Trajectories into Dense Regions. In *ICCV*. IEEE, 1583–1590.

[18] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. 2019. Video Object Segmentation using Space-Time Memory Networks. In *ICCV*. 9226–9235.

[19] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2018. Spatial as deep: Spatial cnn for traffic scene understanding. In *AAAI*, Vol. 32.

[20] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 2016. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *CVPR*. 724–732.

[21] Zequn Qin, Huanyu Wang, and Xi Li. 2020. Ultra fast structure-aware deep lane detection. In *ECCV*. 276–291.

[22] Hongje Seong, Junhyuk Hyun, and Euntai Kim. 2020. Kernelized Memory Network for Video Object Segmentation. In *ECCV*. 629–645.

[23] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. 2018. Pyramid Dilated Deeper ConvLSTM for Video Salient Object Detection. In *ECCV*. 715–731.

[24] Jinming Su, Chao Chen, Ke Zhang, Junfeng Luo, Xiaoming Wei, and Xiaolin Wei. 2021. Structure Guided Lane Detection. *IJCAI* (2021).

[25] Tsung-Ying Sun, Shang-Jeng Tsai, and Vincent Chan. 2006. HSI Color Model Based Lane-Marking Detection. In *ITSC*. IEEE, 1168–1172.

[26] Lucas Tabelini, Rodrigo Berriel, Thiago M Paixao, Claudine Badue, Alberto F De Souza, and Thiago Oliveira-Santos. 2021. Keep your eyes on the lane: Real-time attention-guided lane detection. In *CVPR*. 294–302.

[27] Lucas Tabelini, Rodrigo Berriel, Thiago M Paixao, Claudine Badue, Alberto F De Souza, and Thiago Oliveira-Santos. 2021. Polylanenet: Lane estimation via deep polynomial regression. In *ICPR*. IEEE, 6150–6156.

[28] Jigang Tang, Songbin Li, and Peng Liu. 2021. A review of lane detection methods based on deep learning. *Pattern Recognition* 111 (2021), 107623.

[29] Ziwei Liu Stella X. Yu Tsung-Wei Ke, Jyh-Jing Hwang. 2018. Adaptive Affinity Fields for Semantic Segmentation. In *ECCV*. Springer, 605–621.

[30] TuSimple. 2017. http://benchmark.tusimple.ai.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *NeurIPS* 30 (2017).

[32] Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques, and Xavier Giro-i Nieto. 2019. RVOS: End-to-End Recurrent Network for Video Object Segmentation. In *CVPR*. 5277–5286.

[33] Jun Wang, Tao Mei, Bin Kong, and Hu Wei. 2014. An Approach of Lane Detection Based on Inverse Perspective Mapping. In *ITSC*. IEEE, 35–38.

[34] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. 2021. End-to-End Video Instance Segmentation with Transformers. In *CVPR*. 8741–9750.

[35] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, and Wenxiu Sun. 2021. Efficient Regional Memory Network for Video Object Segmentation. In *CVPR*. 1286–1295.

[36] Yizhuo Zhang, Zhirong Wu, Houwen Peng, and Stephen Lin. 2020. A Transductive Approach for Video Object Segmentation. In *CVPR*. 6949–6958.

[37] Yujun Zhang, Lei Zhu, Wei Feng, Huazhu Fu, Mingqian Wang, Qingxia Li, Cheng Li, and Song Wang. 2021. VIL-100: A New Dataset and A Baseline Model for Video Instance Lane Detection. In *ICCV*. 15681–15690.

[38] Tu Zheng, Hao Fang, Yi Zhang, Wenjian Tang, Zheng Yang, Haifeng Liu, and Deng Cai. 2020. Resa: Recurrent feature-shift aggregator for lane detection. *AAAI* 5, 7 (2020).

[39] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. 2020. Motion-Attentive Transition for Zero-Shot Video Object Segmentation. In *AAAI*, Vol. 34. 13066–13073.