2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Multi-Task Visual Perception with Temporal Feature Fusion for Autonomous Driving

Huei-Yung Lin[1] and Shih-Han Wei[2]

Abstract— **With the rapid developments of autonomous driving technologies, accurate scene perception has become essential for safe and efficient navigation. The key perception tasks such as lane detection, semantic segmentation of road markings and road area, and object detection directly impact vehicle decision-making and obstacle avoidance. However, most existing methods are trained on a single-task dataset, limiting data diversity and reducing performance in complex scenarios or under occlusion and illumination variation. In this work we propose a multi-task perception network based on image sequence input, integrating lane detection, road marking and road area segmentation, and object detection into a unified framework. The network model employs multi-task learning to share features and improve the computational efficiency, and adopts the cross-dataset training paradigm to enhance generalization across tasks. Furthermore, the temporal information from adjacent frames is leveraged to compensate visual degradation in current frames. Experimental results obtained on multiple datasets demonstrate the proposed technique achieves competitive performance compared to state-of-the-art approaches. Code is available at `https://github.com/hank890121/MTVP`**

I. Introduction

Autonomous driving has become one of the most promising application of intelligent system development, where real time and accurate perception of the surrounding environment is crucial to ensuring safe and rational decision-making. The modern autonomous vehicles typically rely on camera-based visual perception for scene understanding. It generally offers rich semantic information at relatively low cost compared to LiDAR or radar-based approaches. These perception systems provide critical input for downstream modules such as localization, route planning, vehicle control, obstacle avoidance, and accident prevention.

In real-world driving environments, it is necessary to have the sensing system deal with multiple visual perception tasks concurrently [1], [2]. Among them, three core capabilities are widely recognized: (1) traffic object detection, which enables the recognition and localization of vehicles, pedestrians, and traffic lights and signs; (2) semantic segmentation of drivable areas, which helps identify navigable road regions for path planning; (3) lane detection, which identifies lane boundaries to support the lane-keeping and steering, playing a vital role in safe and stable driving; and (4) road marking segmentation, which performs pixel-level recognition of crosswalks, arrows, and other markings, enhancing the understanding of regulations and navigation cues for HD map construction and intersection reasoning.

While state-of-the-art deep learning models such as YOLO [3], DeepLab V3+ [4], and Polar R-CNN [5] have achieved remarkable performance in different task domains, deploying separate models for individual tasks increases computational cost, memory usage, and system complexity. This has greatly reduced the applicabilities to real-time systems, especially on embedded platforms. To address these challenges, multi-task learning (MTL) now emerges as an efficient alternative. By sharing a common feature extractor, MTL-based frameworks can simultaneously handle multiple tasks, reduce redundant computation, and improve overall generalization through task synergies.

There are several pioneering studies explored MTL in the context of autonomous driving. In [6], YOLOP introduces a unified model which performs object detection, drivable area segmentation, and lane detection using a single-stage model. HybridNets extends this idea with better decoupling between task heads and a stage-wise training strategy [7]. YOLOPX, on the other hand, improves the overall architecture and task-specific feature adaptation scheme [8]. Nevertheless, most of the existing multi-task frameworks only rely on single dataset (typically BDD100K [9]), where the annotations of all tasks come from the same image set. This induces problems such as dataset imbalance, annotation quality variation and limited task-specific diversities. Since lane markings and road signs might be incomplete or inconsistently labeled in general, the performance and generalization will degrade across the tasks. Moreover, road markings such as arrows, crosswalks, dashed lines, and stop lines play the essential roles in defining traffic semantics and supporting HD map construction.

Unlike lane lines, the road markings are often finer, more diverse in the geometry, and subject to frequent occlusion and wear. Accurate road marking segmentation is thus crucial but still remains under-explored in general MTL frameworks. In addition, lane detection is considered as an unsolved problem in complex environments, especially at nighttime, under rain, or with strong shadows. These issues necessitate more robust, and temporally consistent perception strategies employed for autonomous driving tasks.

To address the challenges, we propose a unified multi-task visual perception framework which jointly performs vehicle detection, lane detection, and semantic segmentation of both drivable regions and road markings. Our network is equipped with a homography-guided temporal fusion module that can effectively integrate those features from multiple consecutive

[1]Huei-Yung Lin is with Department of Computer Science and Information Engineering, National Taipei University of Technology, Taipei 106 , Taiwan, and Department of Electrical Engineering, National Chung Cheng University, Chiayi 621, Taiwan `lin@ntut.edu.tw`

[2]Shih-Han Wei is with Department of Computer Science and Information Engineering, National Taipei University of Technology, Taipei, Taiwan `112598062@cc.ntut.edu.tw`

frames. This design has successfully enhanced the robustness of lane and road marking detection without requiring camera poses or additional sensors. The contributions of this work are as follows:

- We present a multi-task perception model that integrates four tasks in a unified architecture. By sharing feature representations and inference pipeline, the network effectively improves overall computational efficiency and resource utilization.
- A feature fusion module based on consecutive frames is designed to improve the prediction stability of lane line and road marking segmentation tasks.
- A cross-dataset training paradigm is introduced. It addresses the limitation of existing multi-task perception models that relied on a single dataset, not only expanding available training resources but also enabling each sub-task to achieve better performance.

II. Related Work

Multi-Task Learning (MTL) has emerged as an effective paradigm for visual perception of autonomous driving, aiming to jointly perform tasks such as object and lane detection, road marking and drivable area segmentation, etc. In contrast to training multiple single-task networks, MTL significantly reduces the computational cost and memory usage by sharing a common encoder, while enhancing generalization through task synergy. In current literature, several camera-based MTL frameworks have been proposed to deal with driving-related vision tasks in a unified model. YOLOP [6] introduced a real-time multi-task network based on YOLOv4-tiny [10], jointly performing object detection, lane detection, and drivable area segmentation. HybridNets improved the architecture design and training strategy by incorporating BiFPN and stage-wise optimization, at a cost of reduced the inference speed [7]. In [8], YOLOPX proposed an anchor-free decoupled detection head and a lightweight lane head to enhance the flexibility and accuracy of multi-tasking.

Current techniques for multi-task visual perception mostly rely on a single dataset for network training. Since the multi-label annotations can only be provided from the same image set, the information for lane lines and road markings may be sparse, inconsistent, or incomplete, and the diversity of each sub-task is constrained. To mitigate this issue, the cross-dataset training is explored for multitask learning. Kapidis *et al.* proposed a multi-dataset, multitask approach for various egocentric vision tasks, and demonstrated the improvements over single-dataset baselines [11]. Zhou *et al.* [12] presented a face analysis technique based on multitask learning. They utilized cross-dataset learning with each dataset covered one or many task labels. For the multitask road scene perception, PETRv2 [13] adopted the OpenLane [14] and nuScenes [15] datasets in the cross-dataset training pipeline, and performed 3D detection and lane estimation tasks.

Object detection is a core function of autonomous driving systems, responsible for locating and classifying key traffic participants such as vehicles, pedestrians, and traffic signals. Among one-stage detection methods, the YOLO family has progressed from the early anchor-based designs toward more efficient and lightweight architectures. YOLOv7 introduced the E-ELAN backbone with a one-to-many label assignment strategy, which achieves a strong trade-off between detection accuracy and real-time performance [16]. Their subsequent versions continued to refine the design: YOLOv8 enhanced feature aggregation through a C2f module, while YOLOv10 adopted a one-to-one detection paradigm that eliminates the need for non-maximum suppression, making it particularly suitable for multi-task applications [17], [3]. Meanwhile, the Transformer-based detectors have advanced rapidly. DETR established an end-to-end detection framework using bipartite matching [18], which was later improved by variants such as Deformable DETR and DINO to enhance the convergence speed and accuracy [19], [20]. Despite the strong performance, the Transformer-based models typically require more complex training procedures and higher computational costs, so that YOLO-style detectors remain more practical for real-time autonomous driving scenarios.

Lane detection has played an essential role in lane keeping, motion planning, and safe highway driving. Existing studies explore a wide range of methodological paradigms. Semantic segmentation-based approaches, such as SCNN [21], utilize the spatial message passing to produce dense lane predictions, but they often incur high computational overhead and remain sensitive to occlusions and poor visibility conditions. Row-wise or column-wise classification methods, including UFLD, CondLaneNet, and E2E-LMD [22], [23], [24], improve efficiency by predicting lane positions along predefined image slices for reducing output complexity while preserving competitive accuracy. Geometric regression approaches, such as PolyLaneNet and LSTR, focus on learning structured lane representations through polynomial modeling or transformer-based sequence prediction, which enhances geometric consistency [25], [26] . Detection-style frameworks like CLRNet and Polar R-CNN instead formulate lanes as structured detection targets using anchor-based or polar-coordinate representations, allowing seamless integration with object detection pipelines and reducing dependence on post-processing steps like NMS [27], [5]. In addition, temporal modeling methods, exemplified by HomoFusion, further strengthen robustness in the challenging scenarios by aligning and aggregating multi-frame features through homography-based fusion [28].

Road markings, in contrast to lane lines, exhibit a greater variability in shape, are typically small in scale, and are more susceptible to environmental disturbances such as occlusion, motion blur, and pavement degradation. The properties demand perception models with strong high-resolution feature representations, precise geometric sensitivity, and robust semantic interpretation capabilities. Despite their importance, many existing multi-task learning (MTL) frameworks either give limited attention to this task or merge it into the coarse drivable-area segmentation, which results in the loss of fine-grained semantic information. Recent research has attempted to enhance road-marking segmentation through the strategies such as LiDAR-vision fusion and knowledge distillation [29]. However, effectively capturing diversity and complexity of

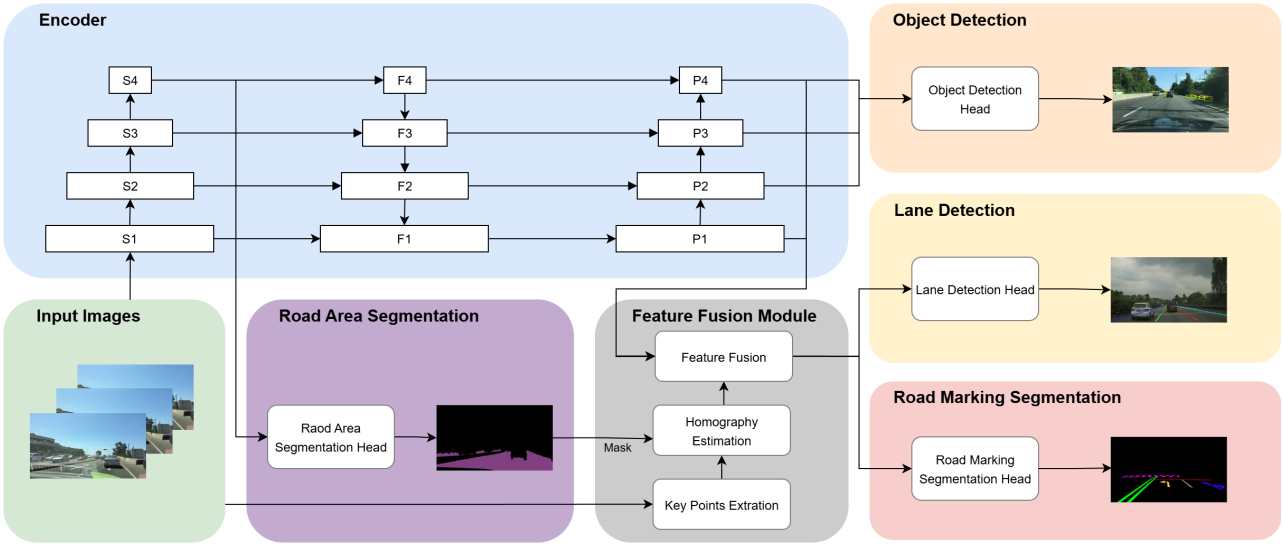


Fig. 1: The proposed multi-task visual perception framework is composed of five key components: object detection, lane detection, road marking segmentation, drivable area segmentation, and a temporal feature fusion module. During training, images are drawn from task-specific datasets and routed only to the modules relevant to their available annotations. In contrast, during inference, each input image is simultaneously processed by all modules to produce comprehensive perception results.

road markings still relies on large-scale, dedicated datasets specifically designed for this purpose [30], [31].

III. Method

Fig. 1 illustrates the overall design of the proposed multi-task cross-dataset perception framework. The system comprises five primary modules: object detection, road marking segmentation, drivable area segmentation, lane detection, and temporal feature fusion. The fusion module is designed to aggregate the semantic cues from consecutive frames, so to improve the consistency and robustness of segmentation outputs. Given a sequence of input images, a shared backbone first extracts visual features, which are then distributed to the respective task-specific heads for prediction.

A. Encoder

Within our multi-task learning framework, the backbone serves as the core component for generating shared feature representations across all tasks. It must effectively preserve low-level geometric details to support structurally sensitive tasks such as lane and road marking detection, and capturing high-level semantic information required for object detection and drivable area segmentation simultaneously. To meet these requirements, we employ PVTv2-B0 as the backbone due to its lightweight pyramid architecture and spatial-reduction attention (SRA) mechanism [32], which lowers computational overhead while maintaining the strong performance in dense prediction scenarios. Its inherent multi-scale feature outputs are well suited for handling the object-centric and pixel-level tasks, making it appropriate for real-time ADAS deployment. Moreover, we incorporate FPN [33] and PAN [34] to enable bidirectional feature fusion across scales. This combination preserves the semantic richness and fine spatial details, and produces robust fused feature maps for the downstream task-specific heads.

B. Decoder

1) Object Detection: For the object detection component, we adopt YOLOv10 as the detection head. In this model, YOLO introduces a Consistent Dual Assignment mechanism that integrates both one-to-many and one-to-one label assignment strategies during training. The one-to-many branch uses a Task-Aligned Assigner (TAL) to allocate multiple positive samples to each object, to provide richer supervisory signals and accelerating convergence. Meanwhile, the one-to-one branch employs top-1 matching to generate a single prediction for each ground-truth, which directly serves as the inference pathway. This dual-branch design enables end-to-end detection without the need for non-maximum suppression, effectively reducing inference latency. To maintain the stable optimization across both of the branches, a consistent matching metric is applied, jointly accounting for classification confidence, IoU-based localization quality, and spatial priors. This unified matching criterion can align the training objectives of the two branches, leading to improved stability and better generalization performance.

2) Lane Detection: We adopt Polar R-CNN [5] as the lane detection module. It is a two-stage, anchor-free framework formulated in polar coordinate space. The architecture consists of three key components: a local polar module (LPM), a global polar module (GPM), and a triplet prediction head. Specifically, the LPM receives the fused feature map P_2' as input and generates polar-form local anchor representations, including the regression features $\boldsymbol{F}^{reg}$ and the corresponding

confidence heatmap $\boldsymbol{F}^{cls}$:

$$\begin{aligned} \boldsymbol{F}^{reg} &\equiv \left\{\theta_j, r_j^l\right\}_{j=1}^{H^l \times W^l} \\ \boldsymbol{F}^{cls} &\equiv \left\{s_j^l\right\}_{j=1}^{H^l \times W^l} \end{aligned} \tag{1}$$

where (θ_j, r_j^l) is the polar representation of lane anchor. The training labels are generated by computing the direction and distance from a local pole to the nearest point of a lane curve. Finally, the set of local poles is given by $\boldsymbol{F}^{\text{cls}} \left\{\hat{s}_j^l\right\}_{j=1}^{H^l \times W^l}$, where $\hat{s}_j^l = 1$ if the j-th local pole is a positive sample, and $\hat{s}_j^l = 0$ otherwise. We use binary cross-entropy (BCE) as the loss function $\mathcal{L}_{\text{cls}}^l$ and smooth L_1 loss $\mathcal{L}_{\text{reg}}^l$ for regression

The GPM receives the anchor proposals and feature maps generated by the LPM. Using ROI Pooling, it samples the fused feature map P_2' to obtain refined lane anchor features, which are subsequently fed into separate classification and regression branches. Conventional lane detection approaches typically rely on a one-to-many prediction paradigm, which often requires Non-Maximum Suppression (NMS) to remove the duplicate results. In contrast, the ternary head design in Polar R-CNN incorporates an additional one-to-one classification branch alongside the one-to-many branch, enabling the model to produce unique lane predictions directly without relying on NMS. Its loss contains both the classification and regression components:

$$\begin{aligned} \mathcal{L}_{\text{cls}}^g &= w_{\text{cls}}^{\text{O2M}} \mathcal{L}_{\text{cls}}^{\text{O2M}} + w_{\text{cls}}^{\text{O2O}} \mathcal{L}_{\text{cls}}^{\text{O2O}} \\ \mathcal{L}_{\text{reg}}^g &= w_{\text{GIoU}}^{\text{O2M}} \mathcal{L}_{\text{GIoU}}^{\text{O2M}} + w_{\text{end}}^{\text{O2M}} \mathcal{L}_{\text{end}}^{\text{O2M}} \end{aligned} \tag{2}$$

where $w_{\text{cls}}^{\text{O2M}}$, $w_{\text{cls}}^{\text{O2O}}$, $w_{\text{GIoU}}^{\text{O2M}}$, and $w_{\text{end}}^{\text{O2M}}$ are the weighting for different loss terms. We use focal loss for $\mathcal{L}_{\text{cls}}^{\text{O2M}}$ and $\mathcal{L}_{\text{cls}}^{\text{O2O}}$, and $\mathcal{L}_{\text{end}}^{\text{O2M}}$ is the smooth L_1 loss from the endpoints of lane lines. Finally, the total loss for lane detection is

$$\mathcal{L}_{lane} = \mathcal{L}_{\text{cls}}^l + \mathcal{L}_{\text{reg}}^l + \mathcal{L}_{\text{cls}}^g + \mathcal{L}_{\text{reg}}^g \tag{3}$$

3) Segmentation: For both drivable area and road-marking segmentation, we employ SegFormer [35] as the core framework. This Transformer-based architecture offers a favorable trade-off between accuracy and computational efficiency. In our network model, we replace the original MiT backbone by PVTv2 to enhance performance while keeping the model size and computational overhead comparable. During the training, we apply a cross-entropy loss to supervise the segmentation outputs, enabling reliable pixel-level classification.

4) Feature Fusion Module: The temporal feature fusion module is designed to improve the perception of static road scenes, such as lane line and marking. It begins by extracting keypoints from consecutive frames using LightGlue [36]. To maintain geometric reliability, extracted keypoints are filtered with a drivable-area segmentation mask so that only road-surface features are preserved. A homography matrix is then computed using the Direct Linear Transform (DLT) method and further refined with RANSAC [37] to eliminate outliers. After the alignment, a pixel-to-pixel attention mechanism is applied to fuse homography-warped features from previous frames with those of the current image.

Given current and previous frame features $P^t[p]$ (queries) and $P^{t-i}[p]$ (keys), a similarity is derived after L_2 normalization

$$a_i = \frac{P^t[p]}{|P^t[p]|_2} \cdot \frac{P^{t-i}[p]}{|P^{t-i}[p]|_2}, \quad i = 1, 2 \tag{4}$$

and the weight is calculated by softmax

$$W_i = \frac{\exp a_i}{\sum_i \exp a_i} \tag{5}$$

Finally, the fused representation is obtained as

$$P'[p^t] = P^t[p] + \sum_i W_i P^{t-i}[p], \quad i = 1, 2 \tag{6}$$

By integrating semantic masking with geometric alignment, the feature fusion module effectively removes dynamic or irrelevant regions, such as vehicles, pedestrians, and surrounding structures, while improving the alignment of road-surface features. As a result, lane markings and structural details are more consistently maintained across frames, producing more stable and reliable representations for downstream tasks, including lane detection and overall road scene understanding.

IV. Experiments

A. Dataset and Implementation

We adopt the BDD100K dataset [9] to evaluate the object detection and road surface segmentation performance. It is a large-scale benchmark containing over 100,000 driving video frames captured under the diverse scenes, weather conditions, and lighting environments. BDD100K provides annotations for multiple perception tasks, including the object detection, semantic segmentation, drivable area and lane marking detection. In this work, we primarily utilize the object detection and semantic segmentation subsets, while also leveraging the drivable area and lane marking annotations to ensure the fair comparisons with existing multi-task learning methods.

For the lane detection evaluation, we adopt the VIL-100 dataset [38], which consists of 100 driving videos and 10,000 high-resolution images (1920×1080). It includes the detailed annotations for ten types of lane markings, including solid lines, dashed lines, curved lanes, and double lines, captured under diverse and challenging conditions, including heavy traffic, shadows, low illumination, and worn/degraded road markings. Designed to assess both robustness and generalization, this dataset provides a comprehensive benchmark for lane detection in complex real-world driving environments.

To assess the road marking segmentation performance, we employ the SeRM dataset [30], a large-scale benchmark developed for semantic segmentation and HD map construction using monocular imagery. SeRM contains 25,158 pixel-level annotated images, 19,998 for training and 5,160 for testing, each with image resolution of 1280×672, collected across diverse traffic scenes in South Korea, including urban roads, intersections, and highways. It provides high-quality, consistent annotations for 17 classes of road markings and traffic signs, making it a dependable benchmark for segmentation evaluation. Following the dataset setting, only the lower half

TABLE I: The experimental results on traffic object detection. It is evaluated with recall rate and mAP@50.

Type	Method	Recall (%)	mAP@50 (%)
Single-task	YOLOv10s [3]	<u>93.3</u>	83.0
Multi-task	YOLOP [6]	89.2	76.5
	HybridNets [7]	92.8	77.3
	YOLOPv2 [39]	91.1	**83.4**
	A-YOLOM (s) [40]	86.9	81.1
	YOLOPX [8]	**93.7**	<u>83.3</u>
	Ours	93.1	<u>83.3</u>

TABLE II: The experimental results on lane detection. It is evaluated with F1@50 and mIoU.

Type	Method	F1@50 (%)	mIoU (%)
Video-Based	MMA-net [30]	83.9	70.5
	RVLD [41]	92.4	**78.7**
	OMR [42]	**93.6**	77.4
Image-Based	ADNet [43]	92.0	<u>78.1</u>
	Polar R-CNN [5]	92.3	77.8
	Ours	<u>92.6</u>	77.9

of each image is annotated, and all evaluations are conducted accordingly.

We use mean average precision (mAP) with mAP@0.5 : IoU ≥ 0.5 and recall as the evaluation metrics of the object detection task carried out on BDD100K. Lane detection on the VIL-100 dataset is assessed using precision, recall, F1-score, and accuracy, where the predicted lanes are considered correct when their overlap with ground truth satisfies an IoU threshold of at least 0.5. For the semantic segmentation tasks, including drivable area and road marking segmentation, we use mean Intersection over Union (mIoU) as the evaluation metric to measure network performance for comparisons.

All experiments are performed on a workstation featuring an NVIDIA GeForce RTX 4090 GPU with 24 GB memory, an Intel i7-13700K processor, and 64 GB system RAM. Input images are resized to 640×384 to achieve a balance between computational efficiency and feature map detail. The models are trained utilizing AdamW as the optimizer with an initial learning rate of 0.0006, a weight decay of 0.001, and a batch size of 8, while a cosine annealing scheduler is employed to ensure stable convergence. A two-stage training scheme is adopted: The branch of drivable area segmentation is trained independently for 30 epochs first; The segmentation decoder is then frozen while the object detection and lane detection branches are trained for an additional 80 epochs, using the segmentation masks used to support feature fusion.

The overall training objective is a weighted sum of task-specific losses:

$$\mathcal{L}_{total} = \lambda_{obj}\mathcal{L}_{obj} + \lambda_{lane}\mathcal{L}_{lane} + \lambda_{rmseg}\mathcal{L}_{rmseg} + \lambda_{raseg}\mathcal{L}_{raseg} \quad (7)$$

where the weights λ_{obj}, λ_{lane}, λ_{rmseg} and λ_{raseg} correspond to object detection, lane detection, road marking segmentation and road area segmentation, and are given by 1.0, 0.7, 0.8 and 0.3, respectively, considering dataset scale and task convergence speed. For lane detection, the number of polar anchors per image is capped at 20 to maintain computational efficiency. For temporal fusion, features from three consecutive frames are aggregated, enhancing the model's ability to capture stable road markings and lane structures over time.

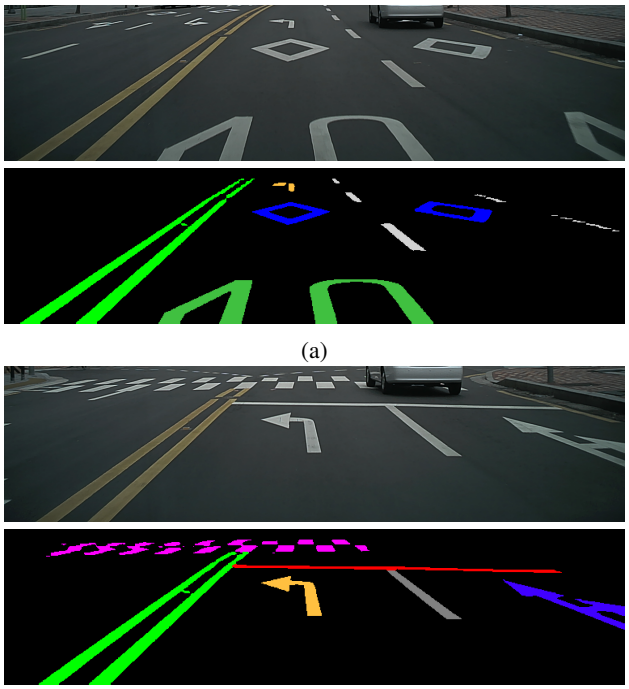
B. Performance Comparison

The results of traffic object detection are presented in Table I. Compared with single-task detectors such as Faster R-CNN and YOLOv5s, all multi-task approaches achieve clear gains in both recall and mAP@50, highlighting the advantages of shared representations and joint optimization. Among multi-task baselines, YOLOPv2 has obtained the highest mAP@50 at 83.4%, while YOLOPX achieves the best recall of 93.7%. Our proposed method delivers a well-balanced performance, attaining a recall of 93.1% and an mAP@50 of 83.3%, which is comparable to YOLOPX. These results have indicated that the proposed framework preserves strong detection accuracy while improving object coverage, demonstrating robust performance in complex driving environments.

The quantitative results for lane detection are summarized in Table II. Among the video-based methods, OMR achieves the highest F1@50 score of 93.6%, whereas RVLD obtains the top mIoU of 78.7%. For image-based methods, ADNet and Polar R-CNN also demonstrate strong performance, both reporting F1 scores exceeding 92%. Our proposed framework attains an F1@50 of 92.6% and an mIoU of 77.9%, placing second overall among the compared methods. Notably, even within a unified multi-task setting, the model has remained competitive with most video- and image-based approaches, indicating its effectiveness in preserving the fine lane details and maintaining the accurate localization without relying on dedicated temporal modeling or separate decoders.

The performance evaluation of road marking segmentation is presented in Table III. Among four compared techniques, DeepLabV3+ achieves the highest mIoU of 79.6%, followed closely by our proposed method at 79.4%, which outperforms both Segformer-B0 and ERFNet. Although DeepLabV3+ is slightly better than ours in terms of the mIoU, it comes with significantly higher computational overhead. Our method can maintain the near-optimal segmentation accuracy while operating within a lightweight and unified multi-task framework. It demonstrates the effectiveness and efficiency for real-time applications. Examples of the road marking segmentation are shown in Fig. 2.

Table IV shows the evaluation result of road area segmentation. DeepLabV3+ achieves the best performance with an mIoU of 79.6%, followed by SFNet-Lite [45] with the mIoU of 77.1%. Segformer-B0 also performs competitively with a score of 76.9%. Our proposed method reaches the mIoU of 76.7%, slightly lower than Segformer-B0 and SFNet-Lite. It is, however, important to note that our model operates within a multi-task setting, without dedicated decoders for road area



(a)

(b)

Fig. 2: The road marking segmentation results obtained from our multi-task perception framework with temporal fusion.

TABLE III: The experimental results on road marking segmentation evaluated with the mIoU.

Method	mIoU (%)
Segformer-B0 [35]	75.9
ERFnet [44]	75.5
DeepLabV3+ [4]	79.0
Ours	**79.4**

segmentation. This demonstrates that our unified framework can still achieve competitive accuracy across tasks, and offer a fairly balanced trade-off between the model complexity and performance.

Table V presents a comparative analysis of several multi-task perception frameworks across the object detection (recall and mAP@50), drivable area segmentation (in DA IoU), and lane line segmentation (in LL IoU). The proposed technique demonstrates strong and well-balanced performance across all evaluated tasks. In object detection, our model achieves a recall of 93.1% and an mAP@50 of 83.3%. While YOLOPX slightly outperforms with recall rate of 93.7% and an mAP of 83.4%, our method maintains comparable accuracy while delivering more balanced overall capability across multiple tasks. For drivable area segmentation, the proposed network attains an IoU of 93.6%, surpassing most existing methods, including YOLOP and HybridNets. Although the gains over YOLOPX and YOLOPv2 are relatively small, the results do reflect the model's robustness in preserving spatial structure and scene consistency. In lane line segmentation, our framework has reached an IoU of 30.9%, significantly outperforming YOLOP and YOLOPv2. Although HybridNets achieves slightly higher lane line accuracy, our approach offers more stable and consistent performance when considering all tasks jointly.

TABLE IV: The experimental results on road area segmentation evaluated with the mIoU.

Method	mIoU (%)
Segformer-B0 [35]	76.9
ERFnet [44]	75.5
DeepLabV3+ [4]	**79.6**
SFNet-Lite [45]	77.1
Ours	76.7

TABLE V: The performance comparison of current state-of-the-art multi-task networks across all different tasks.

Model	Recall	mAP@50	DA IoU	LL IoU
YOLOP [6]	89.2	76.5	91.5	26.2
HybridNets [7]	92.8	77.3	90.5	**31.6**
YOLOPv2 [39]	91.1	**83.4**	93.2	27.2
A-YOLOM [40]	86.9	81.1	91.0	28.8
YOLOPX [8]	**93.7**	83.3	93.2	27.2
Ours	93.1	83.3	**93.6**	30.9

V. Conclusions

In this paper, we present a multi-task perception technique for autonomous driving which simultaneously handles object detection, lane line detection, and road marking and drivable area segmentation. To improve the temporal consistency and feature alignment across tasks, a temporal fusion module is incorporated to utilize information from consecutive frames during inference. In the experimental evaluation on multiple datasets, it shows that our approach outperforms the current state-of-the-art single-task and multi-task baselines. Notably, the framework achieves a well-balanced performance across both detection and segmentation tasks while remaining computationally efficient and lightweight.

Acknowledgment

The support of this work in part by National Science and Technology Council of Taiwan under Grant 114-2221-E-027-096-MY3 is gratefully acknowledged.

References

[1] J. Phillips, J. Martinez, I. A. Bârsan, S. Casas, A. Sadat, and R. Urtasun, "Deep multi-task learning for joint localization, perception, and prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4679–4689.

[2] X. Liang, X. Liang, and H. Xu, "Multi-task perception for autonomous driving," in *Autonomous Driving Perception: Fundamentals and Applications*. Springer, 2023, pp. 281–321.

[3] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, *et al.*, "Yolov10: Real-time end-to-end object detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 107 984–108 011, 2024.

[4] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[5] S. Wang, J. Liu, X. Cao, Z. Song, and K. Sun, "Polar r-cnn: End-to-end lane detection with fewer anchors," *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[6] D. Wu, M.-W. Liao, W.-T. Zhang, X.-G. Wang, X. Bai, W.-Q. Cheng, and W.-Y. Liu, "Yolop: You only look once for panoptic driving perception," *Machine Intelligence Research*, vol. 19, no. 6, pp. 550–562, 2022.

[7] V. Dat, N. Bao, and P. Hung, "Hybridnets: End-to-end perception network," *Pattern Recognition and Image Analysis*, vol. 35, no. 2, pp. 106–118, 2025.

[8] J. Zhan, Y. Luo, C. Guo, Y. Wu, J. Meng, and J. Liu, "Yolopx: Anchor-free multi-task learning network for panoptic driving perception," *Pattern Recognition*, vol. 148, p. 110152, 2024.

[9] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[10] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[11] G. Kapidis, R. Poppe, and R. C. Veltkamp, "Multi-dataset, multitask learning of egocentric vision tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 6618–6630, 2021.

[12] C. Zhou, R. Zhi, and X. Hu, "Cross-dataset face analysis based on multi-task learning," *Applied Intelligence*, vol. 53, no. 10, pp. 12 971–12 984, 2023.

[13] Y. Liu, J. Yan, F. Jia, S. Li, A. Gao, T. Wang, and X. Zhang, "Petrv2: A unified framework for 3d perception from multi-camera images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3262–3272.

[14] L. Chen, C. Sima, Y. Li, Z. Zheng, J. Xu, X. Geng, H. Li, C. He, J. Shi, Y. Qiao, *et al.*, "Persformer: 3d lane detection via perspective transformer and the openlane benchmark," in *European Conference on Computer Vision*. Springer, 2022, pp. 550–567.

[15] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[16] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 7464–7475.

[17] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023.

[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[19] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[20] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv preprint arXiv:2203.03605*, 2022.

[21] X. Pan, J. Shi, P. Luo, X. Wang, and X. Tang, "Spatial as deep: Spatial cnn for traffic scene understanding," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[22] Z. Qin, P. Zhang, and X. Li, "Ultra fast deep lane detection with hybrid anchor driven ordinal classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 5, pp. 2555–2568, 2022.

[23] L. Liu, X. Chen, S. Zhu, and P. Tan, "Condlanenet: a top-to-down lane detection framework based on conditional convolution," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3773–3782.

[24] S. Yoo, H. S. Lee, H. Myeong, S. Yun, H. Park, J. Cho, and D. H. 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*, 2020, pp. 1006–1007.

[25] L. Tabelini, R. Berriel, T. M. Paixao, C. Badue, A. F. De Souza, and T. Oliveira-Santos, "Polylanenet: Lane estimation via deep polynomial regression," in *2020 25th international conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 6150–6156.

[26] R. Liu, Z. Yuan, T. Liu, and Z. Xiong, "End-to-end lane shape prediction with transformers," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3694–3702.

[27] T. Zheng, Y. Huang, Y. Liu, W. Tang, Z. Yang, D. Cai, and X. He, "Clrnet: Cross layer refinement network for lane detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 898–907.

[28] S. Wang, C. Nguyen, J. Liu, K. Zhang, W. Luo, Y. Zhang, S. Muthu, F. A. Maken, and H. Li, "Homography guided temporal fusion for road line and marking segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1075–1085.

[29] R. Yin, Y. Cheng, H. Wu, Y. Song, B. Yu, and R. Niu, "Fusionlane: Multi-sensor fusion for lane marking semantic segmentation using deep neural networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 1543–1553, 2020.

[30] W. Jang, J. Hyun, J. An, M. Cho, and E. Kim, "A lane-level road marking map using a monocular camera," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 1, pp. 187–204, 2021.

[31] H.-C. Hsiao, Y.-C. Cai, H.-Y. Lin, W.-C. Chiu, and C.-T. Chan, "Rlmd: A dataset for road marking segmentation," in *2023 International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan)*. IEEE, 2023, pp. 427–428.

[32] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational visual media*, vol. 8, no. 3, pp. 415–424, 2022.

[33] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[34] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

[35] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[36] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 17 627–17 638.

[37] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[38] Y. Zhang, L. Zhu, W. Feng, H. Fu, M. Wang, Q. Li, C. Li, and S. 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*, 2021, pp. 15 681–15 690.

[39] C. Han, Q. Zhao, S. Zhang, Y. Chen, Z. Zhang, and J. Yuan, "Yolopv2: Better, faster, stronger for panoptic driving perception," *arXiv preprint arXiv:2208.11434*, 2022.

[40] J. Wang, Q. J. Wu, and N. Zhang, "You only look at once for real-time and generic multi-task," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 9, pp. 12 625–12 637, 2024.

[41] D. Jin, D. Kim, and C.-S. Kim, "Recursive video lane detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8473–8482.

[42] D. Jin and C.-S. Kim, "Omr: Occlusion-aware memory-based refinement for video lane detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 129–145.

[43] L. Xiao, X. Li, S. Yang, and W. Yang, "Adnet: Lane shape prediction via anchor decomposition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6404–6413.

[44] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2017.

[45] X. Li, J. Zhang, Y. Yang, G. Cheng, K. Yang, Y. Tong, and D. Tao, "Sfnet: Faster and accurate semantic segmentation via semantic flow," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 466–489, 2024.