Dashcam-Based Ego-Vehicle Speed Estimation via Lane-Aware Spatiotemporal Learning

1st Woong-Chan Byun
CCS Graduate School of Mobility
Korea Advanced Institute of
Science and Technology (KAIST)
Daejeon, Rep. of Korea, 34051
woongchan.byun@kaist.ac.kr

2nd Seung-Hyun Song
Graduate School of
Advanced Security Science and Technology
Korea Advanced Institute of
Science and Technology (KAIST)
Daejeon, Rep. of Korea, 34051
shyun@kaist.ac.kr

3rd Chan-Bin Lim
CCS Graduate School of Mobility
Korea Advanced Institute of
Science and Technology (KAIST)
Daejeon, Rep. of Korea, 34051
chanbin.lim@kaist.ac.kr

4th Dong-Hee Paek
Mechanical Engineering Research Institute
Korea Advanced Institute of
Science and Technology (KAIST)
Daejeon, Rep. of Korea, 34051
donghee.paek@kaist.ac.kr

5th Seung-Hyun Kong*
CCS Graduate School of Mobility
Korea Advanced Institute of
Science and Technology (KAIST)
Daejeon, Rep. of Korea, 34051
skong@kaist.ac.kr

***Abstract*—Dashcams have become widely adopted and are utilized as evidentiary data for traffic accident analysis and collision reconstruction. In particular, ego-vehicle speed is a critical variable in accident investigations, as it is directly related to braking-distance analysis and speeding assessment. However, estimating ego-vehicle speed from dashcam video requires converting the motion observed in the image plane into the real-world domain. This conversion process relies on accurate intrinsic and extrinsic camera parameters. Such requirements are difficult to satisfy for commercial dashcams, which typically exhibit significant lens distortion, do not provide accessible calibration parameters, and are installed with user-specific positions and orientations. To address these limitations, this paper proposes a spatiotemporal learning framework that estimates ego-vehicle speed using only monocular dashcam video, without relying on any geometric camera information. The proposed framework integrates a ResNet-based feature extractor with a ConvLSTM module to model temporal motion patterns. In addition, lane segmentation is incorporated as an auxiliary task to provide geometric priors associated with lane structure and road scale. Experimental results on real-world driving datasets demonstrate that the proposed method reduces RMSE by 22.1% compared with state-of-the-art approaches, while requiring no camera calibration parameters.**

***Index Terms*—Dashcam, Ego-Vehicle Speed Estimation, Lane Segmentation**

I. Introduction

With the rapid increase in the diffusion of vehicular dashboard cameras (dashcams), They have become utilized as key evidence material to determine negligence and reconstruct accident scenarios in the event of traffic collisions [1]. In particular, ego-vehicle speed is an essential element in the accident analysis process for identifying causal relationships, such as calculating collision energy, analyzing braking distance, and determining speeding violations.

Conventionally, vehicle speed is measured via Global Positioning System (GPS) signals. However, in GPS-denied environments—such as tunnels, underpasses, and urban canyons densely populated with skyscrapers signals are frequently interrupted, or data becomes unreliable due to multipath error [2]. Consequently, vision-based speed estimation technology, which calculates vehicle speed by analyzing dashcam video alone without reliance on GPS, has emerged as a critical research topic.

To estimate accurate vehicle speed from video, a process of converting the 2D pixel coordinate system to the 3D world coordinate system is required. This necessitates precise knowledge of intrinsic parameters and extrinsic parameters. However, given the vast variety of dashcams on the market and the idiosyncratic installation positions and angles chosen by users, it is practically difficult or impossible to retrospectively determine the exact parameters of a specific camera during post-accident analysis.

To overcome these limitations, this paper proposes a deep learning-based methodology capable of precisely estimating ego-speed using only visual cues within the image, without any camera calibration information. Specifically, we focus on the fact that lanes on the road contain consistent standardized specifications and geometric information. Lanes provide powerful geometric priors for identifying vanishing points and understanding road perspective, serving as crucial clues for inferring scale even in the absence of camera parameters.

Therefore, we propose an end-to-end network that learns spatiotemporal features of video sequences via a Convolutional LSTM (ConvLSTM) [3]-based recurrent neural network, while

*Corresponding author

simultaneously using lane segmentation as an auxiliary task to significantly improve speed estimation accuracy.

The main contributions of this paper are as follows:

- We present a methodology capable of precise ego-speed estimation by leveraging the geometric context provided by lanes, even in scenarios where intrinsic and extrinsic camera parameters are entirely unknown.
- We proposed a multi-task framework that incorporates lane segmentation as an auxiliary task to explicitly capture the spatiotemporal dynamics of the road.
- To the best of our knowledge, the proposed model achieves a performance improvement of approximately 22.1% in RMSE and 27.3% in MAE compared to current state-of-the-art (SOTA) models in the field of ego-speed estimation.

The paper is organized as follows: Section 2 reviews existing methods for ego-vehicle speed estimation. Section 3 introduces the proposed architecture designed to learn spatiotemporal information based on lanes. Section 4 presents the experimental setup and evaluation results, and Section 5 concludes the paper.

II. Related Works

Research on vision-based ego-speed estimation has evolved from geometric modeling to deep learning-based End-to-End learning, keeping pace with advancements in computer vision technology.

Initially, research primarily focused on utilizing geometric principles. Representatively, Hayakawa et al. [4] proposed a framework integrating deep neural networks to estimate 3D Bounding Boxes, Depth, and Optical Flow, but this method inherently necessitates the camera's Projection and Calibration Matrices to reconstruct 3D world coordinates. To mitigate these parameter dependencies, recent studies have emerged that regress speed directly from video using deep learning. 3DCMA [5] improved performance by incorporating a Lane Segmentation Mask to focus on road regions, while FlexiNet [6] introduced an adaptive feature synthesis network to capture subtle spatiotemporal variations and adapt to diverse driving environments.

However, these state-of-the-art studies rely on benchmark datasets such as KITTI [7] or nuImages [8], which provide precise parameters and geometrically rectified images. In contrast, commercial dashcams frequently exhibit radial distortion due to wide-angle lenses, and obtaining calibration parameters is often impossible. Since existing models are trained under ideal pinhole camera conditions, their performance is not guaranteed on real-world sequences with distortion and unknown parameters. Therefore, this paper proposes a novel methodology capable of robustly estimating speed by leveraging road context information, such as lanes, even in the absence of calibration data and in the presence of image distortion.

III. Proposed Method

In this study, we propose an end-to-end deep learning framework that estimates precise ego-speed by perceiving lane information from monocular dashcam video. The overall network architecture is illustrated in Fig. 1. The proposed model comprises two primary streams: first, a Speed Estimation Branch that extracts spatiotemporal features from the input video sequence and regresses speed; and second, a Lane Segmentation Branch used as an auxiliary task to encourage the model to explicitly learn the geometric structure of the road. The input data consists of a video sequence $X = \{x_1, x_2, \ldots, x_T\}$ composed of T consecutive frames. Each frame passes through a shared ResNet Backbone [9] to be converted into high-dimensional feature maps $FM = \{FM_1, FM_2, \ldots, FM_T\}$. The extracted feature maps FM then branch into the two streams for speed estimation and lane segmentation, respectively.

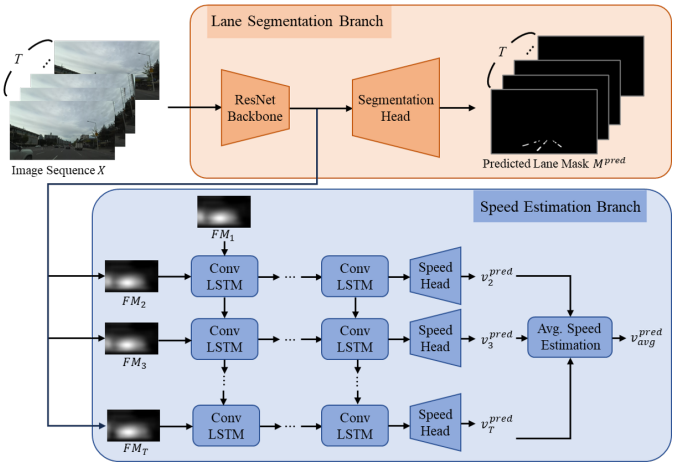

Fig. 1. Overview of the proposed end-to-end network architecture. The model takes an image sequence as input and extracts features via a shared ResNet Backbone. The extracted features are then processed by two parallel branches: the Lane Segmentation Branch (top), which serves as an auxiliary task to learn geometric context, and the Speed Estimation Branch (bottom), which utilizes ConvLSTM for spatiotemporal modeling and weighted average speed regression.

A. Lane Segmentation Branch

To ensure that the shared backbone network clearly learns the Geometric Structure and Perspective information—which are critical cues for speed estimation in monocular images—we utilize the Lane Segmentation Branch as an auxiliary path. This branch is involved in loss calculation only during the training phase and can be optionally omitted during the inference phase. To restore the spatiotemporal feature maps FM, which have reduced spatial resolution after passing through the ResNet backbone, into a precise pixel-level lane mask, the segmentation head performs a series of feature-decoding operations. Specifically, the input FM first passes through a 3×3 convolution layer and a ReLU activation function to adjust channel dimensions and enhance local spatial context. The subsequent 1×1 convolution layer calculates class scores to determine whether each pixel belongs to a Lane or Background. This is then up-sampled to the original input resolution ($H \times W$) via bilinear interpolation to form

the final predicted mask $M^{pred} \in \mathbb{R}^{H \times W}$. This process can be mathematically expressed as follows:

$$M^{pred} = \text{Upsampling}(\text{Conv}_{1\times1}(\text{ReLU}(\text{Conv}_{3\times3}(FM))))$$

During training, this branch is optimized by calculating the Pixel-wise Cross-Entropy Loss between the prediction and the ground truth lane labels. Such Dense Supervision prevents the speed estimation model from relying solely on the global motion of the entire image. Instead, it provides a regularization effect that forces the backbone to explicitly encode structural features decisive for scale perception, such as vanishing points, lane boundaries, and road surface regions.

B. Speed Estimation Branch

The Speed Estimation Branch is responsible for temporally integrating the spatial features extracted from the backbone to accurately estimate both the instantaneous speed per frame v_t^{pred} and the weighted average speed of the entire sequence v_{avg}^{pred}. This branch consists of spatiotemporal modeling, speed regression, and the definition of the loss function for optimization.

1) Spatiotemporal Modeling and Weighted Aggregation: The backbone feature map sequence FM is input into a ConvLSTM [3] network to effectively model the temporal dynamics of the driving scenes. Unlike standard LSTM that flattens inputs into 1D vectors—thereby losing spatial information—ConvLSTM performs all internal gate operations using convolution. This allows it to learn temporal correlations while preserving the inherent spatial structure ($H \times W$) of the feature maps. The hidden state H_t and cell state C_t at each time step t are updated based on the states of the previous time step and the current input as follows:

$$H_t, C_t = \text{ConvLSTM}(FM_t, H_{t-1}, C_{t-1}), \quad t \in \{2, ..., T\}$$

The hidden state H_t, updated via ConvLSTM, contains spatiotemporal context information and is passed to the Speed Head. The Speed Head vectorizes the feature map via Global Average Pooling and then passes it through a Multi-Layer Perceptron (MLP) to regress the scalar speed value v_t^{pred} for every frame.

$$v_t^{pred} = \text{SpeedHead}(H_t)$$

After obtaining the speed for individual frames, we introduce a Linear Weighted Average method instead of a simple average to calculate the final sequence speed v_{avg}^{pred}. This assigns linearly increasing weights to frames closer to the current time step compared to past time steps, thereby sensitively reflecting recent driving trends such as rapid acceleration or deceleration.

$$v_{avg}^{pred} = \frac{\sum_{t=2}^{T} w_t \cdot v_t^{pred}}{\sum_{t=2}^{T} w_t}, \quad w_t \in [1.0, 2.0]$$

2) Multi-task Objective Function: The proposed network employs a Multi-task Loss as the objective function to simultaneously optimize speed estimation precision and lane segmentation accuracy. The total loss function L_{total} is defined as the weighted sum of the frame-wise speed error L_{frame}, the average speed error L_{avg}, and the lane segmentation error L_{seg}.

$$L_{total} = L_{frame} + L_{avg} + \lambda_{seg} L_{seg}$$

L_{frame} is the Mean Squared Error between the predicted speed per frame and the ground truth speed. A masking technique is applied to exclude the first frame from the loss calculation to account for prediction instability in the initial state of the sequence. L_{avg} represents the MSE for the weighted average speed of the entire sequence, while L_{seg} is the aforementioned pixel-wise Cross-Entropy Loss. λ_{seg} is a weighting coefficient that controls the influence of lane segmentation as an auxiliary task on general training.

$$L_{frame} = \frac{1}{T-1} \sum_{t=2}^{T} (v_t^{pred} - v_t^{gt})^2, \quad L_{avg} = (v_{avg}^{pred} - v_{avg}^{gt})^2$$

The rationale for dualizing the speed loss function into frame-level L_{frame} and average-level L_{avg} components is to induce the model to learn both local dynamics and global tendency in a balanced manner. L_{frame} forces the model to follow instantaneous acceleration and deceleration patterns at each time step, preventing the network from simply blurring predictions into an average value. Conversely, L_{avg} acts as a global constraint across the sequence, mitigating temporary prediction noise that may occur in individual frames and securing the stability of the finally calculated average speed value.

IV. Experimental Results

All experiments in this study were conducted on an NVIDIA GeForce RTX 3090 GPU with 24GB of VRAM. For model training, the batch size was set to 1, and the training process was conducted for a total of 40 epochs. We utilized the Adam optimizer for optimization, with the initial learning rate set to 0.0001. The input sequence length T was set to 10, enabling the model to estimate speed based on 10 consecutive video frames.

A. Datasets

The dataset utilized in our experiments was collected from actual urban and suburban roads in Daejeon, South Korea. The data acquisition system captured driving scenes at a resolution of 1280×720 using a front-mounted dashboard camera. Simultaneously, precise ground truth ego-speed data was acquired and synchronized via the vehicle's internal Controller Area Network (CAN) Bus. Specifically, to ensure the model operates robustly across varying driving conditions, we constructed the dataset to encompass diverse road environments, including Suburban, Highway, School Zone, Urban, and Overpass scenarios. Examples of the collected RGB images and their corresponding ground truth lane masks are visualized in Fig. 2. The entire collected dataset was partitioned into training and validation sets; the speed distribution ratio between these sets is presented in Fig. 3. This distribution was rigorously designed to evaluate whether the model can secure generalized performance across low, medium, and high-speed ranges without being biased toward specific speed bands.

Fig. 2. Representative samples from the collected dataset. The dataset covers diverse driving scenarios including Suburban, Highway, School Zone, Urban, and Overpass to ensure environmental diversity. The top row displays the raw RGB images captured by the dashcam, and the bottom row shows the corresponding ground truth lane masks.

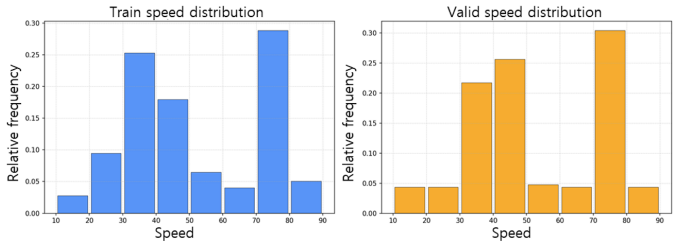

Fig. 3. Distribution of ground truth ego-vehicle speeds in the collected dataset. The histograms display the relative frequency of speed ranges for the training set (top) and the validation set (bottom). The dataset covers a wide range of driving conditions, from low-speed urban driving (10-30 km/h) to high-speed highway driving (over 70 km/h), ensuring the model's generalization capability.

B. Quantitative Results

TABLE I
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART MODELS ON THE COLLECTED DATASET.

Model	**RMSE (km/h)**	**MAE (km/h)**
FlexiNet [6]	16.932	14.724
3DCMA [5]	14.188	10.775
Ours	**11.052**	**7.839**

To evaluate the performance of the model, we utilized Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), which are widely adopted metrics in regression tasks. The definitions of these metrics are as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \quad MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

where N denotes the number of samples, y_i represents the ground truth speed, and $\hat{y}_i$ indicates the predicted speed generated by the model. In our experiments, the weight λ_{seg} for the lane segmentation loss in the proposed model was set to 0.3. Table I presents the quantitative comparison results between the proposed model and state-of-the-art speed estimation models, FlexiNet [6] and 3DCMA [5]. Experimental results demonstrate that the proposed model achieved a significant performance improvement over the comparative models. Specifically, in terms of RMSE, the proposed method reduced the error by approximately 34.7% compared to FlexiNet and 22.1% compared to 3DCMA. Similarly, in terms of MAE, it showed a performance improvement of approximately 46.7% over FlexiNet and 27.2% over 3DCMA. This suggests that learning geometric context through lane segmentation is highly effective for speed estimation, even in the absence of calibration information. Furthermore, across all models, RMSE values were consistently higher than MAE values. This is attributed to the fact that RMSE imposes a quadratic penalty on outliers. It is analyzed that transient prediction errors occurring during rapid acceleration/deceleration phases or in sections with unclear lane markings contributed to the increase in RMSE. However, the proposed model exhibited the smallest gap between RMSE and MAE among the compared models, indicating that it possesses the most stable inference capability even in the presence of outliers.

C. Ablation Study

TABLE II
ABLATION STUDY ON THE IMPACT OF THE AUXILIARY LOSS WEIGHT FOR LANE SEGMENTATION (λ_{seg}).

λ_{seg}	**RMSE (km/h)**	**MAE (km/h)**
0.0 (Baseline)	18.978	16.767
0.1	12.044	8.503
0.3	**11.052**	**7.839**
0.5	14.531	10.568
0.7	14.088	11.017

To analyze the impact of employing lane segmentation as an auxiliary task on speed estimation performance within the proposed multi-task learning framework, we conducted an ablation study by varying the lane segmentation loss weight λ_{seg}. The results are summarized in Table II. First, the Baseline model ($\lambda_{seg} = 0.0$), which utilizes no lane information, yielded the lowest performance, recording an RMSE of 18.978 and an MAE of 16.767. This suggests that relying solely

on learning temporal pixel variations has clear limitations in inferring accurate scale from dashcam where intrinsic and extrinsic parameters are absent. In contrast, when even a small amount of lane segmentation loss was incorporated ($\lambda_{seg} = 0.1$), the RMSE drastically decreased to 12.044, demonstrating a significant improvement in performance. This validates our hypothesis that lane mask information provides strong cues regarding the road's vanishing point and geometric structure, thereby assisting the model in extracting robust features even without calibration data. Notably, the model achieved the best performance when λ_{seg} was set to 0.3, with an RMSE of 11.052 and an MAE of 7.839. However, when the weight was increased beyond 0.3 to 0.5 and 0.7, the RMSE rose again to 14.531 and 14.088, respectively, indicating a degradation in performance. This can be interpreted as the model overfitting to the auxiliary task rather than the main task when the influence of the auxiliary loss becomes excessively large, or the optimization focus becoming diverted, leading to reduced learning efficiency. Consequently, we identified that $\lambda_{seg} = 0.3$ represents the optimal balance, effectively modulating the trade-off between learning geometric features and optimizing speed regression.

V. Conclusion

In this paper, we proposed a ConvLSTM-based multi-task framework for precise ego-speed estimation in dashcam environments. By incorporating lane segmentation as an auxiliary task, the model explicitly learns geometric context to overcome lens distortion and the absence of camera parameters. Experiments demonstrate significant performance improvements over SOTA models, confirming that lane-based geometric cues play a pivotal role in reducing estimation uncertainty. Furthermore, the combined loss function effectively balances instantaneous and global speed learning. This study validates the feasibility of reliable estimation using only monocular cameras. While the current framework establishes a baseline for parameter-free estimation on a single camera type, it serves as a foundational step toward broader applicability. Our future research will focus on scalability across diverse imaging devices. Specifically, we aim to investigate blind parameter estimation techniques to enable precise speed estimation for arbitrary camera models, thereby generalizing the framework to work regardless of camera specifications.

Acknowledgement

This work was supported by the Ministry of Trade, Industry and Resources (MOTIR) grant funded by the Korea government (No. N02250174).

References

[1] E. Giovannini, A. Giorgetti, G. Pelletti, A. Giusti, M. Garagnani, J. P. Pascali, S. Pelotti, and P. Fais, "Importance of dashboard camera (dash cam) analysis in fatal vehicle–pedestrian crash reconstruction," *Forensic Science, Medicine and Pathology*, vol. 17, no. 3, pp. 379–387, 2021.

[2] A. Budiyono, "Principles of gnss, inertial, and multi-sensor integrated navigation systems," *Industrial Robot: An International Journal*, vol. 39, no. 3, 2012.

[3] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[4] J. Hayakawa and B. Dariush, "Ego-motion and surrounding vehicle state estimation using a monocular camera," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 2550–2556.

[5] A. M. Mathew and T. Khalid, "Ego vehicle speed estimation using 3d convolution with masked attention," *arXiv preprint arXiv:2212.05432*, 2022.

[6] A. Ibrahim, K. Kyamakya, and W. Pointner, "Flexinet: An adaptive feature synthesis network for real-time ego vehicle speed estimation," *IEEE Access*, 2025.

[7] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The international journal of robotics research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[8] 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.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.