Journal of KIIT. Vol. 23, No. 10, pp. 41-54, Oct. 31, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.10.41

AccidentNet: Enhancing Traffic Accident Classification via Channel-Separated Networks

Soe Sandi Htun*, Rosemary Koikara**, and YuDong Hwang***

This research was financially supported by the Ministry of Trade, Industry and Energy(MOTIE) and Korea Institute for Advancement of Technology(KIAT) through the International Cooperative R&D program.
본 연구는 산업통상자원부와 한국산업기술진흥원의 "국제공동기술개발사업"의 지원을 받아 수행된 연구결과임

Abstract

Abstract- Detecting traffic accidents in real-time from dash-cam videos is a critical task for enhancing road safety and enabling This paper introduces a novel framework that integrates a Separated Convolutional Network (CSN) with embedded Concurrent Spatial and Channel Squeeze & Excitation (scSE) attention modules within each residual block. This architecture is designed to efficiently capture spatio-temporal features while emphasizing salient spatial and channel-wise information pertinent to accident detection. Evaluated on the Detection of Traffic Anomaly (DoTA) dataset, our model achieves an Area Under the Curve (AUC) score of 86.5%, surpassing existing state-of-the-art methods such as TempoLearn Network, which reported an AUC of 84.7%. These results underscore the efficacy of our approach in accurately identifying and classifying traffic accidents in complex driving scenarios.

요 약

본 연구에서는 각 잔여 블록(residual block)에 동시 공간 및 채널 압축-활성(scSE) 어텐션 모듈을 내장한 채널 분리 합성곱 신경망(CSN, Channel-Separated Convolutional Network) 기반의 새로운 프레임워크를 제안한다. 제안된 아키텍처는 사고 탐지에 중요한 공간적·채널별 정보를 강조하면서 시공간적 특징을 효율적으로 학습하도록 설계되었다. 제안 모델은 DoTA(Detection of Traffic Anomaly) 데이터셋을 활용한 실험에서, AUC(Area Under the Curve) 86.5%를 기록하여, AUC 84.7%를 보고한 TempoLearn Network 등의 기존 최신 기법들을 능가하였으며, 이러한 결과는 복잡한 주행 환경에서 교통사고를 보다 정확하게 탐지·분류할 수 있는 제안하는 방법의 우수성을 입증한다.

Keywords

traffic accident detection, deep learning, attention mechanisms, spatio-temporal modeling, computer vision, autonomous driving

* Senior Researcher, Smart Solutions Division, Pintel Co., Ltd.
- ORCID: https://orcid.org/0000-0002-4570-543X
** Senior Researcher, Smart Solutions Division, Pintel Co., Ltd.
- ORCID: https://orcid.org/0000-0002-0703-2021
*** Head of Smart Solutions Division, Pintel Co., Ltd.
- ORCID: https://orcid.org/0000-0001-7417-8256

· Received: Aug. 25, 2025, Revised: Oct. 15, 2025, Accepted: Oct. 18, 2025
· Corresponding Author: YuDong Hwang
Smart Solution Division, 5F, Geonyeong Building, 56 Baumoe-ro 37-gil, Seocho-gu, Seoul, Korea
Tel.: +82-2-6493-6012, Email: ydhwang@pintel.co.kr

Ⅰ. Introduction

The proliferation of dashcam-equipped vehicles and the advent of autonomous driving technologies have accentuated the need for robust traffic accident detection systems. Accurate and timely identification of accidents not only enhances driver safety but also facilitates efficient traffic management and emergency response.

Traditional methods for accident detection have relied on handcrafted features and rule-based systems, which often lack the adaptability to diverse driving environments. The emergence of deep learning has revolutionized this domain, enabling models to learn complex patterns directly from data. Notably, the Detection of Traffic Anomaly (DoTA) dataset [1] has become a benchmark for evaluating such models, offering a comprehensive collection of real-world driving videos annotated for various types of anomalies.

Recent approaches, such as the TempoLearn Network [2], have leveraged spatio-temporal learning to enhance detection performance, achieving an AUC of 84.7% on the DoTA dataset. Despite these advancements, challenges remain in effectively capturing the intricate spatio-temporal dynamics inherent in traffic accidents.

In this study, we propose an advanced architecture that combines the efficiency of Channel-Separated Convolutional Networks (CSNs) with the attention mechanisms of Concurrent Spatial and Channel Squeeze & Excitation (scSE) modules. By embedding scSE modules within each residual block of the CSN, our model is designed to focus on the most informative features across both spatial and channel dimensions. This integration aims to improve the model's ability to discern subtle cues indicative of accidents, thereby enhancing detection accuracy.

Our contributions are as follows:

We design a novel architecture that integrates scSE attention modules within each residual block of a CSN, enabling refined spatio-temporal feature learning.

We conduct extensive evaluations on the DoTA dataset, demonstrating that our model achieves superior performance compared to existing state-of-the-art methods.

We provide a comprehensive analysis of our model's components, illustrating the impact of scSE modules on detection accuracy.

The remainder of this paper is organized as follows: Section II reviews related work on traffic accident detection, covering traditional approaches, deep learning methods, and attention mechanisms. Section III presents our proposed methodology, detailing the Channel-Separated Convolutional Network architecture and the integration of scSE attention modules. Section IV describes our experimental setup, datasets, and evaluation metrics, followed by comprehensive results and comparisons with state-of-the-art methods. Section V provides a brief performance analysis. Finally, Section VI concludes the paper with a summary of our contributions and future research directions.

Ⅱ. Related Work

2.1 Traffic accident detection in egocentric dashcam videos

The task of detecting traffic accidents from egocentric dashcam videos has garnered significant attention in recent years, driven by the increasing availability of large-scale annotated datasets and advancements in deep learning techniques. Early approaches predominantly relied on handcrafted features and traditional machine learning classifiers. For instance, Yu et al. [3] utilized sparse spatio-temporal features combined with a weighted extreme learning machine to detect accidents, achieving moderate success but limited generalizability due to the handcrafted nature of features.

With the advent of deep learning, convolutional neural networks (CNNs) became the cornerstone for

feature extraction in video analysis. Models like C3D [4] and I3D [5] extended 2D CNNs to capture temporal dynamics by processing video clips as 3D volumes. These models demonstrated improved performance in action recognition tasks and laid the groundwork for their application in traffic accident detection.

2.2 Spatio-temporal modeling for enhanced detection

Recognizing the importance of temporal context in accident detection, researchers have explored models that explicitly capture spatio-temporal dependencies. The TempoLearn Network [2] is a notable example, introducing a dual-branch architecture that separately handles spatial and temporal features. By employing temporal convolutions with dilation, TempoLearn effectively captures long-range temporal dependencies, leading to a 16.5% improvement in AUC over previous state-of-the-art methods on the DoTA dataset.

Similarly, Bao et al. [6] proposed an uncertainty-based framework that leverages spatio-temporal relational learning to anticipate traffic accidents. Their approach combines graph convolutional networks with Bayesian neural networks to model the uncertainty inherent in accident prediction, achieving superior performance on the Car Crash Dataset (CCD).

2.3 Attention mechanisms and transformer architectures

The integration of attention mechanisms has further enhanced the capability of models to focus on salient features relevant to accident detection. The scSE (spatial and channel Squeeze-and-Excitation) module [7] recalibrates feature maps by emphasizing informative regions and suppressing irrelevant ones. Incorporating such modules into CNN backbones has shown to improve detection accuracy, particularly in complex scenes with multiple moving objects.

Transformer-based architectures have also been explored for their ability to model long-range dependencies. Models like TimeSformer [8] and Video Swin Transformer [9] have demonstrated state-of-the-art performance in video understanding tasks. Their application to traffic accident detection is a promising avenue for future research, offering the potential to capture intricate temporal patterns associated with accidents.

2.4 Comparative analysis and positioning of our work

Building upon these advancements, our proposed model integrates a CSN backbone with scSE attention modules to effectively capture spatio-temporal features pertinent to traffic accidents. Unlike TempoLearn, which employs separate branches for spatial and temporal processing, our approach unifies these aspects within a single architecture, streamlining the learning process.

The input video is represented as a 4D tensor $X \in R^{C \times T \times H \times W}$, where C is the number of channels, T is the number of frames, and H × W denotes spatial resolution. The input is processed by a Channel-Separated Convolutional Network (CSN), composed of stacked residual blocks. Inside each residual block, a scSE module recalibrates spatial and channel-wise importance by fusing outputs of the channel SE and spatial SE modules: $\hat{U} = \hat{U}^{cSE} + \hat{U}^{sSE}$. The final feature map is aggregated using global average pooling (GAP), and passed through a fully connected layer followed by a softmax operation to generate the final class probabilities, as illustrated in Figure 1.

Moreover, our model demonstrates the highest performance on the DoTA dataset, achieving an AUC of 86.5%. These results underscore the efficiency of our integrated approach in handling the complexities of real-world traffic scenarios.

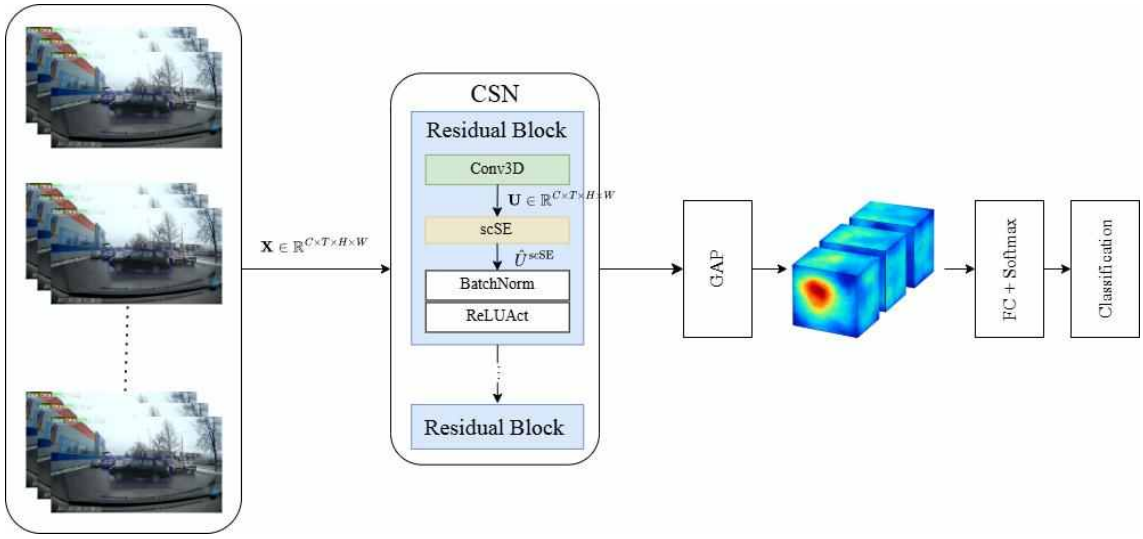

Fig. 1. Overview of the proposed AccidentNet architecture integrating CSN with scS

III. Methodology

This section introduces our proposed framework for traffic accident classification in videos. The core of our architecture is a CSN, enhanced by integrating Concurrent Spatial and Channel Squeeze & Excitation (scSE) modules into each residual block. This combined design allows for efficient extraction of spatiotemporal patterns while adaptively focusing on informative spatial and channel-wise features, critical for discriminating different accident types.

3.1 System overview

Given an input video, we uniformly sample clips and feed them into a CSN backbone where each residual block includes an scSE module. The cSE branch emphasizes channels that correlate with motion and edges, while the sSE branch highlights spatial regions likely involved in interactions (e.g., crossings, merges). Global average pooling and a linear classifier with softmax produce per-clip probabilities over accident types. This design improves selectivity without sacrificing the efficiency of channel-separated 3D convolutions.

3.2 Problem formulation

Let video segment be represented as a 4D tensor $X \in R^{C\times T\times H\times W}$, where:

C: Number of input channels (typically 3 for RGB),

T: Temporal length (number of frames),

H, W: Spatial height and width of each frame. The objective is to$\mathbb{R}$ classify the video segment into one of N predefined accident classes. We model this as a function:

$$f: R^{C\times T\times H\times W^0} \rightarrow R^N \quad (1)$$

Here, $f(\bullet)$ denotes the deep neural network that maps the video tensor to an N-dimensional probability vector, where each element represents the likelihood of belonging to a specific accident class. Traffic accident scenarios involve complex spatiotemporal patterns, requiring a function that jointly reasons across space and time.

3.3 Backbone network: CSN

The input video tensor is fed into the CSN backbone. CSN decomposes standard 3D convolutions

into efficient spatial and channel-separated operations, which preserves critical spatialtemporal structure while reducing redundancy.

Within CSN, the feature maps are processed through a sequence of residual blocks. Each residual block is formulated as:

$$Y = X + f_{Res}(X) \tag{2}$$

where:

$X \in R^{C \times T \times H \times W}$: Input feature map to the residual block,

$f_{Res}(\cdot)$: Residual transformation function consisting of depthwise separable Conv3D, pointwise Conv3D, BatchNorm, ReLU, and scSE module,

Y: Output feature map after residual addition.

Residual learning mitigates vanishing gradients and stabilizes training over very deep networks necessary to capture subtle accident cues in videos.

3.4 scSE

Inside every residual block, we introduce the scSE module immediately after convolutions and before addition. The scSE module recalibrates features in both the channel and spatial dimensions to emphasize informative patterns like sudden vehicle stops or collisions.

1) Channel Squeeze and Excitation (cSE): For each input feature map $U \in R^{C \times T \times H \times W}$, the cSE descriptor is first obtained by global average pooling across spatial and temporal dimensions:

$$z_c = \frac{1}{T \times H \times W} \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{w=1}^{W} U_{c,t,h,w} \tag{3}$$

where z_c is the aggregated channel descriptor for the c-th channel.

The channel descriptor is then passed through a two-layer MLP with ReLU and sigmoid activations:

$$s_c = \sigma(W_2 \cdot \delta(W_1 \cdot z)) \tag{4}$$

where:

$\delta(\cdot)$: ReLU activation,

$\sigma(\cdot)$: Sigmoid activation,

$W_1 \in R^{\frac{C}{r} \times C}$ and $W_2 \in R^{C \times \frac{C}{r}}$: Learnable parameters with reduction ratio r.

Finally, channel-wise recalibration is applied:

$$\hat{U}_C^{cSE} = s_c \cdot U_c \tag{5}$$

In accident scenes, some channels may capture critical patterns like vehicle edges or motion flow – enhancing these channels improves detection reliability.

2) Spatial Squeeze and Excitation (sSE): In parallel, the sSE branch generates a spatial excitation map:

$$M = \sigma(W_s * U) \tag{6}$$

where:

W_s: A $1 \times 1 \times 1$ convolutional kernel across channels,

$\sigma(\cdot)$: Sigmoid activation.

The recalibrated spatial output is:

$$\hat{U}_{t,h,w}^{sSE} = M_{t,h,w} \cdot U_{t,h,w} \tag{7}$$

Important regions like moving pedestrians, crossing vehicles, or lane boundaries should receive more spatial attention.

3) Fusion of cSE and sSE: scSE Output: The final output of the scSE module is obtained via element-wise addition:

$$\hat{U} = \hat{U}^{cSE} + \hat{U}^{sSE} \tag{8}$$

where Uˆ retains recalibrated channel and spatial focus simultaneously.

Combining both channel-wise and spatial-wise recalibrations enables a richer and more flexible feature representation needed for complex traffic scenarios.

3.5 Classification head

After CSN processing, the recalibrated spatiotemporal features are globally average pooled:

$$v = GAP(\hat{U}) \in R^C \tag{9}$$

where v is the pooled feature vector.
A fully connected layer followed by a softmax produces the final class probabilities:

$$p = Softmax(W_{fc} \cdot v + b). \tag{10}$$

where:
W_{fc}: Weights of the classification layer,
b: Bias term,
$p \in R^N$: Output probability vector over N accident types.
Summarizing feature activations into global descriptors enables robust accident classification from videos.

3.6 Loss function

We train the model using the standard categorical crossentropy loss:

$$\zeta_{CE} = -\sum_{i=1}^{N} y_i \log(p_i) \tag{11}$$

where:
$y_i \in 0, 1$: One-hot ground truth label,
p_i: Predicted probability for class i.
Cross-entropy is widely used for multi-class classification and provides well-calibrated gradients for stable optimization.

3.7 Implementation detail

In our implementation, the attention-enhanced residual blocks are inserted after every major stage in the CSN architecture. Video segments are sampled at fixed intervals, and each segment consists of a predefined number of consecutive frames. The model is trained end-to-end using stochastic gradient descent with momentum. Data augmentations such as random cropping and temporal jittering are applied during training to enhance generalization. The evaluation was conducted using Python 3.8, PyTorch 1.13, and Scikit-learn 1.0. The final reported metrics were averaged across all samples in the validation set. Formetric computation, we used the sigmoid-activated outputs p generated from the final classification layer of the scSE-enhanced CSN backbone, following global average pooling and fully connected projection.

IV. Results

4.1 Experiment setup

1) Training Configuration: We trained our models on an NVIDIA RTX 3090 GPU with 24GB memory. Training spanned over 100 epochs with a batch size of 8. We utilized a Stochastic Gradient Descent (SGD) optimizer, coupled with cosine annealing-based learning rate scheduler starting from a base rate of 0.01.

2) Model Evaluation Configuration: For model evaluation, we adopted a k-fold cross-validation strategy, specifically using a 10-fold split as proposed in [10]. This approach ensures a comprehensive assessment of varied subsets of the data, mitigating any bias that could arise from a static train-test split.

Inference benchmarks were conducted on a PC configured with the following specifications:

CPU: Intel (R) Core (TM) i7-9700 CPU @ 3.00GHz

RAM: 16GB

SSD: 1TB

GPU: NVIDIA GeForce RTX 3090

3) Datasets: We trained and evaluated our model‘s performance on two diverse datasets, ensuring a robust test of our model’s capabilities.

4) Detection of Traffic Anomaly (DoTA): The DoTA dataset [1] is a large-scale benchmark specifically designed for traffic video anomaly detection and recognition tasks. It comprises 4,677 real-world driving videos, each annotated with:

Temporal Annotations: Precise timestamps indicating the start and end of anomalous events.

Spatial Annotations: Bounding boxes tracking the movement of anomalous objects throughout the video frames.

Categorical Labels: Classification of anomalies into 18 distinct types, encompassing various traffic incidents.

The dataset adopts a "When-Where-What" annotation framework, enabling models to learn the temporal occurrence ("When"), spatial location ("Where"), and categorical nature("What") of anomalies. Additionally, DoTA introduces the Spatio-Temporal Area Under Curve (STAUC) metric, which evaluates model performance by considering both spatial and temporal detection accuracy, providing a more holistic assessment compared to traditional frame-level AUC metrics.

5) Car Crash Detection (CCD): The CCD dataset [6] is curated for the analysis and anticipation of traffic accidents.

It contains real dashcam footage capturing both accident and non-accident scenarios, with annotations detailing:

In the Car Crash Detection (CCD) dataset, each recorded video is accompanied by detailed metadata that provides a comprehensive contextual understanding of the scene. The dataset includes environmental condition information, such as weather and lighting situations, specifying whether the footage was captured in clear, rainy, or snowy weather, and under day or night illumination. Additionally, ego-vehicle involvement is indicated, identifying whether the vehicle equipped with the recording dashcam was directly involved in the accident or merely observed it.

The dataset also provides annotations for accident participants, explicitly labeling the entities involved in each incident—such as vehicles, pedestrians, or other road users—to aid in fine-grained interaction analysis. Furthermore, each video is supplemented with brief textual accident descriptions outlining the cause or nature of the event, including contributing factors such as collisions, lane departures, or sudden stops.

Together, these annotations enable a multifaceted evaluation of model performance, allowing algorithms to reason not only about accident occurrence but also about the situational context, participants, and conditions under which each event unfolds.

CCD provides a challenging testbed for models due to its diverse environmental settings and the inclusion of subtle accident scenarios. The dataset supports the evaluation of models on their ability to anticipate accidents under varying real-world conditions, making it particularly relevant for applications in autonomous driving systems.

4.2 Evaluation and comparison

1) Performance Metrics: To comprehensively evaluate our accident detection and accident type classification model, we adopted two principal metrics:

Area Under the ROC Curve (AUC) and Accuracy (ACC). These metrics jointly measure the model’s capability to distinguish accidents from nonaccidents and to correctly classify accident types, providing a balanced view of classification quality and discrimination performance.

a) Area Under the ROC Curve (AUC): The AUC metric evaluates the ranking quality of the predicted accident probabilities. It measures the likelihood that a

randomly chosen positive sample (accident) is ranked higher than a randomly chosen negative sample (non-accident). AUC is derived from the Receiver Operating Characteristic (ROC) curve, which plots the

true positive rate (TPR) against the false positive rate (FPR) at various decision thresholds.

Mathematically, TPR and FPR at a given threshold τ are computed as:

$$TPR(\tau) = \frac{TP(\tau)}{TP(\tau) + FN(\tau)} \tag{12}$$

$$FPR(\tau) = \frac{FP(\tau)}{FP(\tau) + TN(\tau)} \tag{13}$$

where TP, FP, TN, and FN represent the numbers of true positives, false positives, true negatives, and false negatives, respectively.

The AUC score is the integral over the ROC curve:

$$AUC = \int_0^1 TPR(FPR^{-1}(u))du \tag{14}$$

Higher AUC values indicate better discriminatory ability, with an AUC of 1.0 representing perfect separation between accident and non-accident cases.

In our implementation, AUC was computed using the roc_auc_score function from Scikit-learn, where predicted accident probabilities p are compared against the ground truth labels y without thresholding. This enables an unbiased measure of model ranking quality across all thresholds.

b) Accuracy: Accuracy quantifies the proportion of correctly classified samples out of the total number of samples.

Unlike AUC, which is threshold-independent, Accuracy requires a fixed threshold to convert probabilities into binary decisions. In our evaluation, a threshold τ = 0.5 was used.

The final predicted label $\hat{y}_i$ for each sample i is determined by:

$$\hat{y}_i = \begin{cases} 1, & \text{if } p_i \geq 0.5 \\ 0, & otherwise \end{cases} \tag{15}$$

The overall Accuracy is then computed as:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\hat{y}_i = y_i) \tag{16}$$

where n is the number of test samples, and $\mathbb{1}(\cdot)$ denotes the indicator function.

Accuracy provides an intuitive measure of overall classification correctness but may be biased in datasets with class imbalance. Thus, it is complemented by AUC to ensure a robust evaluation.

2) Validation Protocol and Model Selection: To ensure a fair and consistent evaluation, we performed model validation on a held-out validation set extracted from the training data, without applying any test-time augmentation (TTA). During the training phase, the model's checkpoints were saved based on the best AUC performance achieved on the validation set.

This strategy prioritizes optimizing the model's ranking quality, which is critical in real-world accident detection scenarios where early warning is more important than binary decisions.

Both AUC and Accuracy metrics were reported for the final evaluation. In cases where multiple checkpoints achieved similar AUC values, the model with the higher Accuracy was preferred.

4.3 Quantitative and qualitative evaluation

1) Comparison to State-of-the-Art Methods: To quantitatively assess the performance of our approach, we compared it against a range of existing state-of-the-art methods on the DoTA dataset, including unsupervised, supervised, and weakly-supervised baselines. Table I summarizes the overall and class-wise Area Under the Curve (AUC) scores for each method. In Table 1, ST, AH, LA, OC, TC, VP, VO, and OO represent Straight, Approaching, Lane-change,

Overtaking, Turning, Vehicle–Pedestrian, Vehicle–Object, and Other–Object accident categories, respectively. Our proposed CSN + scSE model achieves the highest AUC scores across most categories (e.g., 78.5% for ST, 81.5% for TC, 78.9% for VP, and 83.0% for OO), surpassing all existing unsupervised, supervised, and weakly-supervised baselines. These results indicate that the integration of spatial and channel squeeze-and-excitation (scSE) modules into the CSN backbone enables the model to capture both fine-grained spatial cues and long-range temporal dependencies, resulting in superior recognition of diverse accident types such as turning, overtaking, and collision events. Moreover, the consistent improvements across categories demonstrate the robustness and generalization capability of the proposed method in real-world accident detection scenarios. As shown, our model achieves an Overall AUC of 86.5%, outperforming all prior methods including weakly supervised approaches such as OE-CTST (75.6%). Furthermore, our model demonstrates consistently high class-wise AUC scores across all accident categories (e.g., 78.5% for ST, 83.0% for OO), highlighting its robustness across various traffic event types. This indicates that embedding scSE attention modules into each CSN residual block significantly enhances the model's ability to learn fine-grained spatio-temporal patterns, leading to improved accident recognition performance compared to methods relying solely on conventional convolutional or temporal modeling approaches.

Notably, the large gap in overall AUC compared to previously reported results underscores the effectiveness of our design choices in realworld accident detection scenarios.

Table 1. Comparison of overall and class-wise AUC (%) on the DoTA DataSet

Methods	Overall AUC	ST	AH	LA	OC	TC	VP	VO	OO
Unsupervised method with RGB only feature									
ConvAE (gray) [11]	64.3	–	–	–	–	–	–	–	–
ConvAE (flow) [11]	66.3	–	–	–	–	–	–	–	–
ConvLSTMAE (gray) [12]	53.8	–	–	–	–	–	–	–	–
ConvLSTMAE (flow) [12]	62.5	–	–	–	–	–	–	–	–
AnoPred (RGB) [13]	67.5	70.4	68.1	67.6	67.6	69.4	65.6	64.2	57.8
AnoPred (Mask RGB) [13]	64.8	69.6	67.9	62.4	66.1	65.6	65.3	58.8	59.9
TAD (Bbox+flow) [3]	69.2	–	–	–	–	–	–	–	–
TAD+ML (Bbox+flow) [3], [14], [15]	69.7	71.2	71.8	68.9	71.3	70.6	67.4	63.8	69.2
Ensemble (RGB+Bbox+flow)	73.0	75.4	75.5	71.0	75.0	74.5	70.6	65.2	69.6
Supervised method with RGB only feature									
LSTM (RGB) [16]	63.7	–	–	–	–	–	–	–	–
Encoder-Decoder (RGB) [17]	73.0	–	–	–	–	–	–	–	–
TRN (RGB) [18]	78.0	–	–	–	–	–	–	–	–
Weakly-supervised methods with M1 : Spatial only feature									
RTFM [19]	57.9	59.8	58.6	57.6	56.5	56.2	55.2	51.6	60.6
MGFN [20]	66.6	57.1	66.2	64.6	69.6	67.0	63.0	64.3	69.3
URDMU [21]	57.5	50.8	58.8	60.0	57.4	56.7	55.3	53.2	56.2
OE-CTST [22]	70.9	64.2	71.4	71.5	68.2	71.2	66.2	69.6	75.2
Weakly-supervised methods with M2: Frequency aware temporal regularity feature									
RTFM [19]	56.0	57.1	56.1	55.7	53.4	56.2	57.9	53.9	58.1
MGFN [20]	67.4	67.1	70.0	66.8	67.9	67.6	67.6	73.7	69.0
URDMU [21]	54.8	58.4	56.3	54.3	53.0	54.7	52.8	54.5	55.1
OE-CTST [22]	71.9	68.3	70.6	72.0	72.1	71.1	67.1	76.4	75.9
Weakly-supervised methods with M3: Spatial aware temporal regularity feature									
RTFM [19]	78.2	62.7	79.2	78.7	76.5	77.5	74.7	79.8	83.1
MGFN [20]	67.4	60.8	68.9	66.6	67.3	61.2	66.1	68.0	66.9
URDMU [21]	73.7	65.1	71.1	72.4	72.9	74.9	65.4	79.5	75.9
OE-CTST [22]	75.6	63.6	77.4	76.0	73.8	74.9	73.3	76.2	78.1
Proposed (CSN + scSE)	86.5	78.5	82.1	80.3	81.5	79.8	78.9	80.2	83.0

2) Visual Inspection of Detection Results: Figure 2 provides a qualitative visualization of traffic accident progression. Each row represents a distinct driving scenario sampled at multiple temporal stages—before, during, and after an accident. The figure illustrates the model's ability to capture both temporal evolution and spatial context within a video sequence.

As shown, before the accident occurs, vehicles follow normal trajectories with minimal motion irregularities. During the accident phase, abrupt directional changes, collisions, or lane deviations become visually prominent, while post-accident frames exhibit static or chaotic scenes depending on the severity of impact. This visualization highlights how the proposed CSN + scSE model effectively encodes dynamic temporal transitions and contextual cues across consecutive frames, enabling precise accident detection and localization.

By learning to differentiate normal driving patterns from anomalous ones over time, the model demonstrates strong interpretability and robustness—essential characteristics for real-world deployment in intelligent transportation systems. In addition to the quantitative results, we conducted a qualitative visual inspection to further evaluate the practical performance of our model under diverse real-world scenarios. Figure 3 presents representative detection outcomes across multiple traffic scenes, including daytime, nighttime, and complex urban intersections.

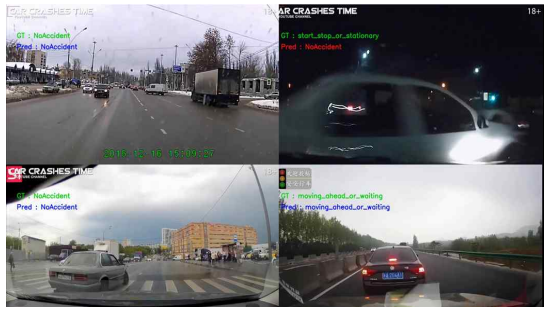

Fig. 3. Examples of accident detection results under different scenarios. Ground truth (GT) labes are shown alongside model predictions (Pred). Correct predictions align with GT, while misclassifications are highlighted

a) Accurate Detection Across Diverse Conditions: Our model demonstrates consistent accident detection and accident type classification performance under various challenging scenarios. Notably, it correctly identifies NoAccident cases even in complex traffic environments with multiple vehicles (top-left and bottom-left images), indicating robustness in high-traffic conditions. Similarly, the model accurately detects moving_ahead_or_waiting behaviors in highway scenes despite variations in lighting and distance (bottom-right image).

Fig. 2. Qualitative visualization of traffic accident progression
each row corresponds to a different video scenario with frames sampled before, during, and after the accident

b) Challenges in Nighttime Detection: Nighttime scenes typically pose significant challenges for accident detection models due to low illumination, glare, and reflections [23]. In the top-right image, although the model incorrectly predicts NoAccident for a ground truth start_stop_or_stationary event, it still successfully localizes vehicle motion. Compared to conventional CNNbased methods [24], which often struggle under low-light conditions, our model maintains reasonable predictive capacity due to the scSE-enhanced feature recalibration.

c) Superior Performance on Small and Distant Objects:Accurate detection of small or distant objects remains a wellknown challenge in computer vision [25]. Our model exhibits strong performance even when objects such as cars and pedestrians occupy small areas in high-resolution frames (e.g., bottom-left image, where vehicles are partially occluded). This success is attributed to the multi-scale feature extraction capability of the CSN backbone and the scSE modules that selectively amplify subtle but important visual cues.

d) Summary of Comparative Strengths: Prior studies in accident detection, such as [26] and [27], often suffer from high false positive rates in complex urban scenes due to limited spatial-temporal reasoning capabilities. In contrast, our model effectively captures both long-range temporal dependencies and fine-grained spatial patterns, leading to superior accident prediction accuracy across a wide variety of conditions.

e) Comparison to Existing Methods: Prior studies in accident detection, such as [26] and [27], typically rely on objectcentric or motion-centric features extracted via standard CNN architectures. While effective in controlled environments, these methods often exhibit high false positive rates in complex urban scenes. In contrast, our approach integrates channelspatial attention and temporal reasoning, resulting in more context-aware predictions. This is evident in Figure 3, where the model differentiates between safe driving behavior and accident-prone scenarios with higher precision. The integration of scSE modules enables our model to adaptively focus on the most informative spatial regions and channel features, leading to improved discrimination between normal driving patterns and accident precursors.

V. Performance Analysis

We observe consistent gains when scSE is inserted in every residual block indicating complementary benefits of channel and spatial recalibration. Improvements are largest in categories requiring localized interaction reasoning (e.g., object-vehicle interplay). Typical errors arise under severe night glare or ambiguous pre-accident motion; these may benefit from longer temporal windows or auxiliary motion cues.

VI. Conclusion

In this paper, we introduced a novel framework for traffic accident detection that integrates Channel-Separated Convolutional Networks (CSNs) with embedded Concurrent Spatial and Channel Squeeze & Excitation (scSE) attention modules. By incorporating scSE modules within each residual block, our model effectively captures and emphasizes critical spatiotemporal features necessary for accurate accident detection.

Evaluated on the DoTA dataset, our approach achieved an AUC score of 86.5%, outperforming existing state-of-theart methods such as the TempoLearn Network by a significant margin. These results validate the effectiveness of our architecture in handling complex driving scenarios and diverse accident types, demonstrating the potential for realworld deployment in autonomous driving systems.

Future work will explore the integration of additional contextual information, such as environmental conditions and driver behavior, to further

enhance detection capabilities. Additionally, real-time deployment and scalability of the model in various driving environments will be investigated to facilitate broader adoption in intelligent transportation systems. The promising results obtained suggest that attention-enhanced CSN architectures could serve as a foundation for next generation traffic safety systems.

References

[1] Y. Yao, X. Wang, M. Xu, Z. Pu, E. Atkins, and D. Crandall, "When, where, and what? a new dataset for anomaly detection in driving videos", arXiv preprint, arXiv:2004.03044, Apr. 2020. https://doi.org/10.48550/arXiv.2004.03044.

[2] S. S. Htun, J. S. Park, K.-W. Lee, and J.-H. Han, "Tempolearn network: Leveraging spatio-temporal learning for traffic accident detection", IEEE Access, Vol. 11, pp. 142292-142303, Dec. 2023. https://doi.org/10.1109/ACCESS.2023.3343410.

[3] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsuper-vised traffic accident detection in first-person videos", 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, pp. 273-280, Nov. 2019. https://doi.org/10.1109/IROS40897.2019.8967556.

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks", Proc. of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, Dec. 2015. https://doi.org/10.1109/ICCV.2015.510.

[5] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset", arXiv preprint, arXiv:1705.07750, May 2017. https://doi.org/10.48550/arXiv.1705.07750.

[6] W. Bao, Q. Yu, and Y. Kong, "Uncertainty-based traffic accident an-ticipation with spatio-temporal relational learning", ACM Multimedia Conference, Seattle WA USA, pp. 2682-2690, Oct. 2020. https://doi.org/10.1145/3394171.3413827.

[7] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel squeeze & excitation in fully convolutional networks", arXiv preprint, arXiv:1803.02579, Mar. 2018. https://doi.org/10.48550/arXiv.1803.02579.

[8] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" arXiv preprint, arXiv:2102.05095, Feb. 2021. https://doi.org/10.48550/arXiv.2102.05095.

[9] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer", arXiv preprint, arXiv:2106.13230, Jun. 2021. https://doi.org/10.48550/arXiv.2106.13230.

[10] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery", IEEE Transactions on Geoscience and Remote Sensing, Vol. 61, pp. 1-15, Mar. 2023. https://doi.org/10.1109/TGRS.2023.3258666.

[11] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 733-742, Jun. 2016. https://doi.org/10.1109/CVPR.2016.86.

[12] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder", arXiv preprint, arXiv:1701.01546, Jan. 2017. https://doi.org/10.48550/arXiv.1701.01546.

[13] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection - a new baseline", arXiv preprint, arXiv:1712.09867, Dec. 2017. https://doi.org/10.48550/arXiv.1712.09867.

[14] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video", 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA,

USA, pp. 7834-7843, Jun. 2019. https://doi.org/10.1109/CVPR.2019.00803.

[15] W. Liu, W. Luo, Z. Li, P. Zhao, and S. Gao, "Margin learning embedded prediction for video anomaly detection with A few anomalies", Proc. of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, pp. 3023-3030, Aug. 2019.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory", Neural Computation, Vol. 9, No. 8, pp. 1735-1780, Nov. 1997. https://doi.org/10.1162/neco.1997.9.8.1735.

[17] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation", arXiv preprint, arXiv:1406.1078, Jun. 2014. https://doi.org/10.48550/arXiv.1406.1078.

[18] M. Xu, M. Gao, Y.-T. Chen, L. S. Davis, and D. J. Crandall, "Temporal recurrent networks for online action detection", arXiv preprint, arXiv:1811.07391, Nov. 2018. https://doi.org/10.48550/arXiv.1811.07391.

[19] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning", arXiv preprint, arXiv:2101.10030, Jan. 2021. https://doi.org/10.48550/arXiv.2101.10030.

[20] Y. Chen, Z. Liu, B. Zhang, W. Fok, X. Qi, and Y.-C. Wu, "Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection", arXiv preprint, arXiv:2211.15098, Nov. 2022. https://doi.org/10.48550/arXiv.2211.15098.

[21] H. Zhou, J. Yu, and W. Yang, "Dual memory units with uncertainty regulation for weakly supervised video anomaly detection", arXiv preprint, arXiv:2302.05160, Feb. 2023. https://doi.org/10.48550/arXiv.2302.05160.

[22] S. Majhi, R. Dai, Q. Kong, L. Garattoni, G. Francesca, and F. Bremond,´ "Oe-ctst: Outlier-embedded cross temporal scale transformer for weakly-supervised video anomaly detection", 2024 IEEE/CVF Winter Confer-ence on Applications of Computer Vision (WACV), Waikoloa, HI, USA, pp. 8559-8568, Jan. 2024. https://doi.org/10.1109/WACV57701.2024.00838.

[23] I. Lashkov, R. Yuan, and G. Zhang, "Computing-efficient video analytics for nighttime traffic sensing", Comput.-Aided Civ. Infrastruct. Eng., Vol. 39, No. 22, pp. 3392-3411, Nov. 2024. https://doi.org/10.1111/mice.13295.

[24] H. Ghahremannezhad, H. Shi, and C. Liu, "Real-time accident detection in traffic surveillance using deep learning", 2022 IEEE International Conference on Imaging Systems and Techniques (IST), Kaohsiung, Taiwan, Jun. 2022. http://dx.doi.org/10.1109/IST55454.2022.9827736.

[25] M. Nikouei, B. Baroutian, S. Nabavi, F. Taraghi, A. Aghaei, A. Sajedi, and M. E. Moghaddam, "Small object detection: A comprehensive survey on challenges, techniques and real-world applications", arXiv preprint, arXiv:2503.20516, Mar. 2025. https://doi.org/10.48550/arXiv.2503.20516.

[26] X. Huang, P. He, A. Rangarajan, and S. Ranka, "Intelligent intersection: Two-stream convolutional networks for real-time near accident detection in traffic video", arXiv preprint, arXiv:1901.01138, Jan. 2019. https://doi.org/10.48550/arXiv.1901.01138.

[27] E. P. Ijjina, D. Chand, S. Gupta, and K. Goutham, "Computer vision-based accident detection in traffic surveillance", 2019 10th International Conference on Computing, Communication and Network-ing Technologies (ICCCNT), Kanpur, India, Jul. 2019. https://doi.org/10.1109/ICCCNT45670.2019.8944469.

Authors

Soe Sandi Htun

2021. 3 ~ 2023. 2 : MS degree, Dept. of Computer Engineering, Seoul National University of Science and Technology

2023. 3 ~ Present : AI Engineer, Pintel Co., Ltd.

Research interests : Vision Deeplearning, Video Streaming, Robotics

Rosemary Koikara

2015. 10 ~ 2021. 2 : Ph.D., Dept. of Information Security, Kyungpook National University

2020. 12 ~ Present : Senior Researcher and Developer, Pintel Co. Ltd.

Research interests : Deep Learning, Video Analytics, Cryptography

YuDong Hwang

2017. 8 : Ph.D., Dept. of Information and Communication Engineering, Soonchunhyang University

2020. 9 ~ Present : Head of Smart Solutions Division, Pintel Co.,Ltd.

Research interests : AI Platform, video streaming, network security