MCSCA: Multi-dimensional Collaborative Spatial-Channel Attention Network for Traffic Sign Recognition

Jiazheng Xu[1], Xiao Huang[2(✉)], Jinlai Zhang[2(✉)], Yuanhao Yang[2]

[1] Elite Engineering School, Changsha University of Science and Technology, Changsha, 410114, Hunan, China
[2] College of Mechanical and Vehicle Engineering, Changsha University of Science and Technology, Changsha, 410114, Hunan, China

Abstract. Traffic sign recognition is a safety-critical perception task in intelligent transportation systems, requiring accurate classification under complex real-world conditions including illumination variation, viewpoint changes, motion blur, and environmental degradation. Existing methods often rely on single-branch attention mechanisms that capture only partial feature dependencies, limiting robustness under degraded visual conditions. To address these limitations, we propose a Multi-dimensional Collaborative Spatial-Channel Attention (MCSCA) network that comprises three complementary attention branches, namely Neuron Saliency Enhancement (NSE), Spatial-Channel Collaborative Calibration (SCC), and Cross-Dimensional Interaction (CDI), through a learnable Softmax-weighted adaptive fusion strategy. The three branches operate in parallel on shared intermediate feature maps, simultaneously enhancing neuron-level saliency, spatial-channel contextual dependency, and cross-dimensional structural interaction. The fused representation is further stabilized via residual connection. The proposed model is built upon a lightweight residual backbone with multi-scale feature aggregation and is trained using AdamW with warmup cosine scheduling, Cut-Mix/Mixup augmentation, and label smoothing. Experiments on GTSRB demonstrate that MCSCA achieves 99.89% validation accuracy, 99.97% precision, and 99.80% recall at 633.4 FPS with only 4.38M parameters, maintaining competitive performance while preserving real-time inference efficiency. Robustness evaluation on GTSRB-C, a corrupted benchmark covering 8 camera corruption types at 5 severity levels, shows a mean corruption accuracy (mCA) of 81.81% and a composite RobScore of 76.91, with near-perfect robustness under photometric corruptions and graceful degradation under additive noise and motion blur. These results validate the effectiveness of the proposed multi-branch collaborative attention design for robust traffic sign recognition under real-world perturbations.

Keywords: Traffic Sign Recognition, Multi-branch Attention, Spatial-Channel Attention, Robustness Evaluation

1 Introduction

Traffic sign recognition plays a crucial role in the efficient operation of modern intelligent transportation systems and in ensuring road safety[4, 14, 16, 12, 20, 6].With the acceleration of urbanization and the rapid development of autonomous driving technology, road environments are becoming increasingly complex and dynamic. How to achieve accurate traffic sign recognition in such complex scenarios has become a major challenge for researchers and engineers alike. Effective traffic sign recognition not only helps enhance the perception capabilities of driver assistance systems but also enables the real-time acquisition of road information to provide early warnings of potential risks (such as speed limit changes or hazardous road sections), thereby significantly improving driving safety. Furthermore, precise traffic sign recognition is a vital component of intelligent transportation systems, capable of enhancing the decision-making capabilities of autonomous driving systems, optimizing traffic flow efficiency, and ultimately improving the traffic environment and the travel experience for residents.

Traffic sign recognition is an intelligent perception technology that utilizes machine vision and artificial intelligence to automatically detect traffic signs on roads and accurately classify and identify their meanings. Since signs in real world street scenes are often subject to disturbances such as small objects, complex backgrounds, varying lighting conditions, and differences in scale, this task is inherently highly challenging. In recent years, with the continuous development of Convolutional Neural Networks (CNNs) and Visual Transformers, deep learning models based on these architectures have been widely applied in traffic sign recognition. With the emergence of models that synergistically compose convolutional and attention mechanisms, researchers have begun combining Transformers with CNNs and attention mechanisms to further improve the accuracy and robustness of traffic sign recognition. In these studies, Transformer networks are used to capture global contextual dependencies, while CNNs and attention modules are employed to enhance the representation of local key features. Currently, such hybrid models have achieved preliminary applications on public traffic sign datasets and in real world road scenes. The German traffic sign recognition benchmark developed by Stallkamp et al. [17] provides a standardized evaluation foundation for traffic sign classification models. Ciregan et al. [2] demonstrated that deep convolutional networks can effectively extract local discriminative features such as color, edges, texture, and shape from traffic sign images, while Zhu et al. [30] further pointed out that traffic signs in real world street scenes often present challenges such as small objects, complex backgrounds, and varying lighting conditions. Addressing the challenges of recognition in complex road scenes, Zhang et al. [28] enhanced the model's adaptability to scale variations through multi-scale feature extraction and cascaded feature fusion. Min et al. [10] combined semantic scene understanding with structured traffic sign localization to mitigate issues of small object missed detection and background misclassification, while An et al. [1] utilized a cascaded attention modulation fusion mechanism to strengthen the representation of key regions. It is evident that research on traffic sign recognition has shifted from early efforts focused on improving single-class classification accuracy to addressing the reliable recognition of small objects, fine grained categories, complex backgrounds, and real world road disturbances.

With the advancement of deep learning models [31-33], traffic sign recognition methods have gradually evolved from traditional CNN architectures toward the synergistic integration of convolutional layers, attention mechanisms, and Transformers. Wang et al. [22] proposed DK-Former, which integrates convolutional random Fourier features with a Transformer architecture, enabling the model to balance the representation of local details with the modeling of global dependencies; Zheng and Jiang [29] and Wang et al. [21] evaluated the performance of visual Transformers in traffic sign classification tasks, noting that while Transformers excel at modeling global relationships, traffic sign recognition remains highly dependent on local symbols, color boundaries, and fine grained structural features. Dosovitskiy et al. [4] introduced the Transformer into image recognition tasks, while Peng et al. [13] further demonstrated the complementary value of local textures and global context by coupling convolutional local features with Transformer global representations. Attention mechanisms also provide significant support for enhancing discriminative features in traffic sign recognition. Yang et al. [24] proposed the NSE parameter-free attention module, Si et al. [15] proposed the SCC spatial-channel collaborative attention mechanism, and Misra et al. [11] proposed the CDI cross-dimensional interaction module. while Woo et al. [23] introduced a joint channel-spatial attention module. This demonstrates that effective features do not exist solely in a single spatial location or a single channel dimension, but are distributed across local textures, spatial structures, channel semantics, and cross-dimensional interactions.

To address these challenges, this paper proposes MCSCA, a Multi-dimensional Collaborative Spatial-Channel Attention network for traffic sign recognition. MCSCA introduces three complementary attention branches, namely Neuron Saliency Enhancement (NSE), Spatial-Channel Collaborative Calibration (SCC), and Cross-Dimensional Interaction (CDI), which operate in parallel on shared intermediate feature maps and fused through a learnable Softmax-normalized adaptive fusion strategy. The model is built upon alightweight residual backbone and trained with AdamW, warmup cosine scheduling, CutMix/Mixup augmentation, and label smoothing. Experiments on GTSRB demonstrate that MCSCA achieves 99.89% accuracy at 633.4 FPS with only 4.38M parameters. Robustness evaluation on GTSRB-C, covering 8 camera corruption types at 5 severity levels, yields a mean corruption accuracy (mCA) of 81.81% and a composite RobScore of 76.91, with near-perfect robustness under photometric corruptions (Fog mCA: 99.85%, Rain mCA: 98.87%) and graceful degradation under additive noise and motion blur.

The contributions of this paper can be summarized as follows:

- We propose MCSCA, a multi-branch collaborative attention module comprising NSE, SCC, and CDI in parallel via learnable Softmax-normalized fusion, producing more discriminative representations for traffic sign recognition.
- We present a robustness-oriented training strategy combining diverse augmentation, CutMix/Mixup regularization, label smoothing, and AdamW with warmup-cosine scheduling, achieving strong robustness on GTSRB-C without corruption-specific training data.

- We construct GTSRB-C, a corrupted benchmark applying 8 corruption types at 5 severity levels to the GTSRB validation split (104,800 images), and define a composite RobScore metric for comprehensive robustness evaluation.

2 Preliminary

In this section, we formulate the traffic sign recognition problem and introduce the attention-based feature representation perspective underlying the proposed framework.

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a labeled traffic sign dataset, where $x_i \in \mathbb{R}^{H\times W\times 3}$ represents an RGB image and $y_i \in \{1,2,\ldots,C\}$ denotes its corresponding category label among C classes. The objective of traffic sign recognition is to learn a parameterized mapping $f_\theta : \mathbb{R}^{H\times W\times 3} \to \{1,\ldots,C\}$ that minimizes the empirical classification loss:

$$\mathcal{L}_{\text{cls}} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \ell(f_\theta(x_i), y_i) \tag{1}$$

where $\ell(\cdot,\cdot)$ denotes the cross-entropy loss with label smoothing.

In real world scenarios, traffic sign images are frequently affected by environmental degradations such as rain, fog, blur, illumination variation, and sensor noise. Let $\tilde{x}_i = \mathcal{C}(x_i, t, s)$ denote a corrupted sample generated by a corruption operator $\mathcal{C}$ of type $t \in \mathcal{T}$ and severity level $s \in \{1,2,3,4,5\}$. The learning objective therefore requires the model to maintain discriminative and robust feature representations on both clean samples $\tilde{x}_i$ and corrupted samples $\tilde{x}_i$ without additional fine-tuning on corrupted data.

Given an intermediate feature representation $F \in \mathbb{R}^{C\times H\times W}$ extracted by a convolutional backbone, attention mechanisms aim to generate a recalibrated representation

$$\mathrm{F}' = \mathcal{A}(\mathrm{F}) \tag{2}$$

where $\mathcal{A}(\cdot)$ denotes an attention transformation that emphasizes informative responses while suppressing irrelevant or noisy features. Existing attention mechanisms mainly focus on different aspects of feature dependency modeling. NSE [24] performs neuron-level importance estimation through an energy-based formulation without introducing additional learnable parameters. SCC [15] jointly models spatial and channel dependencies to enhance contextual feature representation. CDI [11] captures cross-dimensional interactions by computing attention across multiple tensor permutations.

These mechanisms exhibit complementary characteristics in modeling finegrained saliency, spatial-channel correlation, and cross dimensional dependency. Motivated by this observation, the proposed Multi-branch Collaborative Spatial- Channel Attention (MCSCA) framework combines multiple attention perspectives under a unified adaptive fusion strategy to achieve more comprehensive feature recalibration and robust traffic sign recognition.

3 Methodology

In this section, we first provide an overview of the proposed MCSCA framework. Then, we introduce the three core attention components, including the Neuron Saliency Enhancement (NSE) branch, the Spatial-Channel Collaborative Calibration (SCC) branch, and the Cross-Dimensional Interaction (CDI) branch. Next, we describe the learnable feature fusion strategy used to integrate multi-branch attention representations. Finally, we present a corruption-aware robustness optimization strategy for improving recognition stability under degraded visual conditions.

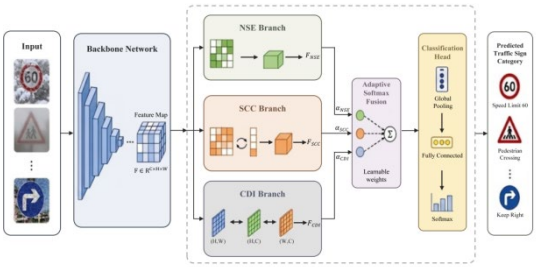

Fig. 1. Overview of the proposed MCSCA-based traffic sign recognition framework.

3.1 Overview of MCSCA

The overall MCSCA-based traffic sign recognition framework is illustrated in Fig. 1. Given an input traffic sign image, the backbone network first performs hierarchical feature extraction and generates an intermediate feature map

$$F \in \mathbb{R}^{C\times H\times W} \tag{3}$$

which encodes both low level visual patterns and high level semantic representations, including color, edge, contour, texture, and category-specific structural information. The extracted feature map is subsequently delivered to the proposed MCSCA module for collaborative multi-branch attention enhancement.

The proposed MCSCA module contains three parallel enhancement branches: the Neuron Saliency Enhancement (NSE) branch, the Spatial-Channel Collaborative Calibration (SCC) branch, and the Cross-Dimensional Interaction (CDI) branch. These branches enhance the shared input feature from three complementary perspectives, namely neuron-level saliency perception, spatial-channel collaborative dependency modeling, and cross-dimensional structural interaction. Their outputs, denoted as F_{nse}, F_{scc}, and F_{cdi}, are then combined through a learnable Softmax-normalized weighted fusion strategy, as detailed in Section 4.2.

The detailed internal structure of the proposed MCSCA module will be further described in Section 4.2 and illustrated in Fig. 2.

Within a residual block, let

$$U = \mathcal{F}_{conv}(X) \tag{4}$$

denote the convolutional feature extracted from the input feature X, where $F_{conv}(\cdot)$ represents the convolutional transformation. The proposed MCSCA module refines the extracted feature as

$$\widetilde{U} = \mathcal{M}_{MCSCA}(U) \tag{5}$$

where $\mathcal{M}_{MCSCA}(\cdot)$ denotes the proposed multi-branch collaborative attention enhancement function.

The output of the residual block is then formulated as

$$Y = \widetilde{U} + \mathcal{S}(X) \tag{6}$$

where $\mathcal{S}(\cdot)$ denotes either an identity shortcut mapping or a 1×1 convolution used for channel dimension alignment.

Finally, the enhanced feature representation is delivered to the classification head for category prediction. During inference, class probabilities are generated through the Softmax function, and the final traffic sign category is determined according to the maximum response probability.

Benefiting from the collaborative multi-branch attention mechanism and adaptive weighted fusion strategy, the proposed MCSCA can simultaneously capture neuron-level saliency, spatial-channel contextual dependency, and cross-dimensional structural interaction. Consequently, the framework produces more robust and discriminative feature representations, making it highly suitable for challenging driving scenarios involving small targets, occlusion, motion blur, il- lumination variation, weather corruption, sensor noise, and complex background interference.

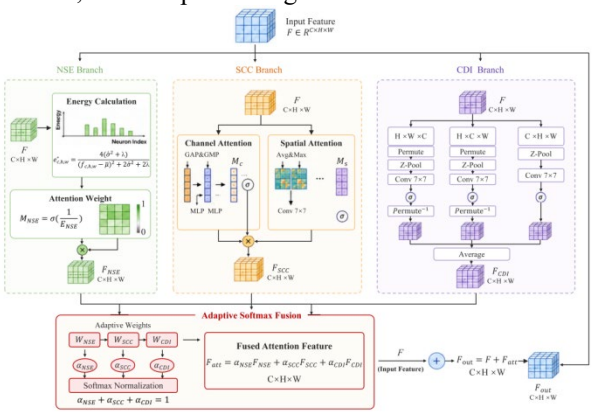

Fig. 2. Detailed structure of the proposed MCSCA module featuring multi-branch parallel collaborative recalibration.

3.2 Multi-branch Collaborative Attention Enhancement

To improve the discriminative representation capability of the proposed framework, we introduce a multi-branch collaborative attention enhancement mechanism during intermediate feature refinement. Since traffic sign characteristics are jointly distributed across local details, semantic channels, spatial regions, and structural dependencies, relying on a single attention mechanism is insufficient for comprehensive feature enhancement under complex driving environments.

Let the intermediate feature map extracted by the backbone network be denoted as

$$F \in \mathbb{R}^{C \times H \times W} \tag{7}$$

where C, H, and W denote the channel number, feature height, and feature width, respectively. To comprehensively enhance the representation capability of F , the proposed framework introduces three complementary attention branches operating in parallel on the shared input feature.

Specifically, the Neuron Saliency Enhancement (NSE) branch focuses on neuron-level importance perception through energy-guided feature recalibration, enabling the model to emphasize informative local responses while sup- pressing irrelevant activations. The Spatial-Channel Collaborative Calibration (SCC) branch jointly models spatial dependency and channel interrelation to enhance contextual semantic consistency across different feature regions. Meanwhile, the Cross-Dimensional Interaction (CDI) branch captures structural correlations across multiple tensor dimensions, thereby improving the representation capability for geometric and structural traffic sign patterns.

Given the shared input feature F , the three collaborative attention branches generate enhanced representations as

$$F_{nse} = \mathcal{M}_{NSE}(F) \tag{8}$$

$$F_{scc} = \mathcal{M}_{SCC}(F) \tag{9}$$

$$F_{cdi} = \mathcal{M}_{CDI}(F) \tag{10}$$

where $\mathcal{M}_{NSE}(\cdot)$, $\mathcal{M}_{SCC}(\cdot)$, $\mathcal{M}_{CDI}(\cdot)$ and denote the neuron saliency enhancement function, spatial-channel collaborative calibration function, and cross-dimensional interaction function, respectively.

Unlike conventional sequential attention stacking, the proposed collaborative architecture allows all branches to independently process the original feature representation in parallel, thereby preserving complementary attention informa- tion from different perspectives while avoiding excessive information coupling. Through this design, the framework can simultaneously refine neuron saliency, contextual dependency, and cross-dimensional structural interaction within a unified feature enhancement process.

3.2.1 Neuron Saliency Enhancement Branch

The first branch of the proposed collaborative attention framework deploys a Neuron Saliency Enhancement (NSE) mechanism to achieve fine-grained, neuron-level feature discrimination. Unlike conventional attention operators that aggregate weights exclusively along isolated spatial or channel vectors, the NSE branch mathematically formalizes a three-dimensional energy paradigm to evaluate the individual visual contribution of each neuron without injecting redundant trainable parameters. This localized parameter-free characteristic is exceptionally suitable for traffic sign recognition tasks, as highly discriminative textual or numerical cues often reside within micro-scale spatial regions embedded in dense channels.

Given the intermediate feature map $F \in \mathbb{R}^{C\times H\times W}$, the NSE branch quantifies the cognitive saliency of each target neuron by measuring its linear separability against surrounding background responses within the same channel cluster. For a specific neuron $f_{c,h,w}$, its objective energy function is defined as:

$$e^*_{c,h,w} = \frac{4(\hat{\sigma}^2 + \lambda)}{(f_{c,h,w} - \hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{11}$$

where $\hat{\mu}$ and $\hat{\sigma}^2$ denote the channel-wise mean and variance of neuron responses, respectively, and λ is a regularization hyperparameter inserted for numerical stability. The analytical minimum of this energy function inversely reflects the distin- guishability of the current coordinate position. A lower energy status implies a higher spatial-channel saliency. Let

$$E_{nse} = \{e^*_{c,h,w}\} \tag{12}$$

represent the full-dimension energy topography computed by the NSE branch. The saliency-calibrated feature representation is then derived via element-wise modulation:

$$F_{nse} = \sigma\left(\frac{1}{E_{nse}}\right) \odot F \tag{13}$$

where $\sigma(\cdot)$ specifies the Sigmoid gating function and $\odot$ denotes element-wise multiplication. By enforcing this parameter-free point-wise evaluation, the framework aggressively crystallizes localized foreground pattern, such as speed limit digits or specific warning symbols, while effectively shielding the representation from global background clutter.

3.2.2 Spatial-Channel Collaborative Calibration Branch

The secondary parallel track incorporates a lightweight Spatial-Channel Collaborative Calibration (SCC) mechanism, designed to jointly modulate semantic channel fields and geographical spatial coordinates in a consolidated pipeline. Within traffic sign representations, channel dimensions inherently dictate abstract global attributes (e.g., color profiles and global texture layouts), whereas spatial planes govern geometric boundary definitions and coordinate configurations. To optimize this dual-domain dependency, the SCC branch sequentially orchestrates channel-wise semantic amplification and spatial-wise coordinate refinement.

Given the incoming shared feature map F, global average pooling and global maximum pooling operations are initiated simultaneously to generate two mutually complementary channel descriptors. These descriptors are subsequently projected through a shared multi-layer linear mapping function $f_{mlp}(\cdot)$, imple- mented via 1×1 convolutions, to derive the contextual channel attention vector:

$$M_c(F) = \sigma\big(f_{mlp}(\mathrm{AvgPool}(F)) + f_{mlp}(\mathrm{MaxPool}(F))\big) \tag{14}$$

where $\sigma(\cdot)$ signifies the Sigmoid activation. The channel-calibrated intermediate representation is explicitly formulated as:

$$F_c = F \odot M_c(F) \tag{15}$$

Conditioned on the channel-recalibrated matrix F_c, the spatial gate is further constructed by computing cross-channel descriptive statistics. The channel-wise mean and maximum planes are densely concatenated and mapped through a large-kernel spatial convolution layer to formalize the structural spatial gate:

$$M_s(F_c) = \sigma\big(\mathrm{Conv}_{7\times7}([\mathrm{Avg}_c(F_c), \mathrm{Max}_c(F_c)])\big) \tag{16}$$

where $\mathrm{Avg}_c(\cdot)$ and $\mathrm{Max}_c(\cdot)$ define the channel-wise average and maximum projection operations, respectively. The final modulated output of the SCC track is expressed as:

$$F_{scc} = F_c \odot M_s(F_c) \tag{17}$$

This serial co-dependency modeling ensures that macro-scale properties, such as the geometric borders of signs (e.g., triangles, octagons) and color boundaries, are reinforced against severe environmental illumination variations and partial background occlusions.

3.2.3 Cross-Dimensional Interaction Branch

The tertiary branch introduces a Cross-Dimensional Interaction (CDI) mechanism to capture coupled structural and spatial dependencies across multiple non-traditional tensor axes. In complex driving scenarios, category differences are heavily correlated with geometric orientation, contour topology, and viewpoint variations, which are easily corrupted by camera tilt or perspective distortion. Because conventional pooling paradigms compress entire dimensions and inevitably cause severe information bottlenecking, the CDI branch explicitly constructs cross-dimensional interaction flows across the channel, height, and width boundaries via mathematical tensor permutation.

For the baseline input feature F, the CDI branch executes three parallel sub-tracks to formalize cross-dimensional linkages. To extract rich statistical descriptors without parameter inflation, a Z-pooling operator is defined by concate- nating channel-wise maximum and average pooling profiles along the designated target axis:

$$Z(F) = [\mathrm{MaxPool}_c(F), \mathrm{AvgPool}_c(F)] \tag{18}$$

The cross-dimensional attention activation mapping is then generalized as:

$$G(F) = F \odot \sigma\big(\mathrm{BN}(\mathrm{Conv}_{7\times7}(Z(F)))\big) \tag{19}$$

where $\mathrm{BN}(\cdot)$ represents batch normalization and $\sigma(\cdot)$ represents the Sigmoid function.

To facilitate direct communication between orthogonal tensor dimensions, permutation operations are deployed to rotate the coordinate axes prior to attention mapping. Let $\mathcal{P}_{cw}(\cdot))and(\mathcal{P}_{ch}(\cdot)$ denote the tensor permutation operators for channel-width and channel-height interaction modeling, respectively. The interaction responses generated across the rotated tensor views are expressed as:

$$F_{cw} = \mathcal{P}_{cw}^{-1}\big(G(\mathcal{P}_{cw}(F))\big) \tag{20}$$

$$F_{ch} = \mathcal{P}_{ch}^{-1}\big(G(\mathcal{P}_{ch}(F))\big) \tag{21}$$

$$F_{hw} = G(F) \tag{22}$$

Ultimately, the complete output of the CDI branch is obtained by symmetrically averaging the three dimensionally interacted feature representations:

$$F_{cdi} = \frac{1}{3}(F_{cw} + F_{ch} + F_{hw}) \tag{23}$$

By aggregating structural correlations across all coordinate viewpoints, the CDI branch injects rotational and scaling invariance into the model, ensuring highly stable geometric representations under viewpoint distortions and motion blur.

3.2.4 Adaptive Multi-branch Fusion

The proposed framework extracts three complementary feature representations from the Neuron Saliency Enhancement (NSE), Spatial-Channel Collaborative Calibration

(SCC), and Cross-Dimensional Interaction (CDI) branches, denoted as F_{nse}, F_{scc}, and F_{cdi}, respectively. These branches capture local saliency, spatial-channel context, and cross-dimensional structural interactions.

Although all three branches belong to the attention family, they are designed with distinct functional emphases. The NSE branch enhances neuron-wise saliency for fine-grained local discrimination, the SCC branch models spatial-channel dependencies for contextual consistency, and the CDI branch captures cross-dimensional interactions to strengthen structural representation. Together, they provide complementary feature modeling across local, contextual, and structural perspectives.

However, since the contributions of different branches are inherently unequal, simple summation or concatenation is insufficient to model their relative importance and may introduce redundant or suboptimal feature responses. To address this issue, we introduce an adaptive multi-branch fusion strategy that learns branch importance weights in an end-to-end manner.

To achieve weighted combination at the feature level, we explicitly define the learnable branch weight vector as:

$$\boldsymbol{w} = [w_{nse}, w_{scc}, w_{cdi}] \tag{24}$$

where w_{nse}, w_{scc}, and w_{cdi} denote the backpropagation-trainable importance parameters corresponding to the NSE, SCC, and CDI branches, respectively.

To ensure comparability among branch contributions and satisfy the normalization constraint, the Softmax function is applied to normalize the learnable branch weights:

$$[\alpha_{nse}, \alpha_{scc}, \alpha_{cdi}] = \mathrm{Softmax}([w_{nse}, w_{scc}, w_{cdi}]) \tag{25}$$

where α_{nse}, α_{scc}, and α_{cdi} denote the normalized fusion coefficients of the three collaborative attention branches. The normalized coefficients strictly satisfy

$$\alpha_{nse} + \alpha_{scc} + \alpha_{cdi} = 1 \tag{26}$$

Based on the learned adaptive coefficients, the collaboratively enhanced feature representation is formulated as

$$F_{att} = \alpha_{nse}F_{nse} + \alpha_{scc}F_{scc} + \alpha_{cdi}F_{cdi} \tag{27}$$

To preserve the integrity of the original feature distribution and stabilize optimization during training, a residual connection is further introduced between the fused attention representation and the original intermediate feature:

$$F_{out} = F + F_{att} \tag{28}$$

where F_{out} denotes the final output feature of the proposed collaborative attention enhancement module.

Through the proposed adaptive fusion strategy, the framework can automatically learn globally optimized contribution weights for different attention branches according to their relative importance during training. Consequently, the fused feature representation simultaneously incorporates neuron-level local saliency, spatial-channel contextual dependency, and cross-dimensional structural interaction.

Unlike fixed fusion strategies such as summation or concatenation, the proposed MCSCA adopts learnable Softmax-based weighting to adaptively balance the contribution of each branch, improving feature complementarity while reducing redundancy.

3.3 Robustness-Oriented Training Strategy

To improve model generalization and robustness under diverse visual degradations, we adopt a training strategy that combines diverse data augmentation, mixed-sample regularization, and stable optimization scheduling. Rather than relying on explicit corruption-aware loss formulations, the proposed strategy implicitly improves robustness by broadening the training distribution to cover appearance variations that overlap with common camera corruptions.

Data Augmentation. Training images undergo a series of geometric and photometric augmentations: random resizing to 40×40 followed by random cropping to 32×32, random rotation ($\pm 20°$), random affine transformations (translation $\pm 10\%$, scale 0.85-1.15, shear 5°), colour jitter (brightness 0.4, contrast 0.4, saturation 0.3, hue 0.05), random grayscale (p=0.05), and random erasing (p=0.2, scale 0.02-0.1). Among these, colour jitter implicitly simulates the global luminance and contrast shifts introduced by fog and rain, contributing to the near-perfect photometric robustness observed in Section 5.4. Random affine transformations provide partial coverage of viewpoint distortion and mild motion blur patterns, while random erasing simulates partial occlusion of sign regions.

Mixed-sample Regularization. CutMix [25] and Mixup [27] are applied with equal probability (50% each) during the first 40 training epochs using mixing coefficient $\lambda \sim \mathrm{Beta}(0.3,0.3)$. These strategies encourage the model to learn interpolation-consistent and spatially distributed feature representations, reduc- ing over-reliance on local discriminative cues and improving generalization to unseen appearance variations. The alternating application of region-level replace- ment (CutMix) and linear interpolation (Mixup) provides complementary regu- larization effects, jointly improving the model's tolerance to partial corruption and occlusion of sign regions.

Optimization and Scheduling. The optimizer is AdamW with an initial learning rate of 10^{-3} and weight decay of 2×10^{-4}. A warmup-cosine annealing schedule is adopted, consisting of a 5-epoch linear warmup followed by cosine decay to a minimum learning rate factor of 10^{-2}. Gradient clipping with maximum norm 5.0 is applied throughout training to prevent instability. Cross-entropy loss with label smoothing $\varepsilon = 0.1$ is used as the training objective, providing soft supervision that improves calibration and reduces overconfidence under distributional shift. All models are trained for 60 epochs with mixed-precision (AMP) enabled on a single NVIDIA GPU.

The combination of these strategies yields a model that generalises well to corrupted inputs despite being trained exclusively on clean images. As demonstrated in the robustness experiments (Section5.4), the training pipeline provides strong implicit coverage for photometric corruptions, while additive noise and heavy occlusion (snow) represent out-of-distribution challenges that remain open for future improvement.

4 Experimental Results

4.1 Datasets

We evaluate the proposed model on two benchmarks: clean recognition on GTSRB and corruption robustness on GTSRB-C.

GTSRB. A large-scale benchmark [18] of 51,839 real-world traffic sign images across 43 categories, covering illumination changes, viewpoint variations, motion blur, and partial occlusion. The training split is divided 80:20 (seed 42) into 31,367 training and 7,842 validation samples; the test split contains 12,630 images.

GTSRB-C. A corrupted variant of GTSRB constructed by applying 8 camera corruption types-snow, fog, rain, Gaussian noise, impulse noise, uniform noise, front-back motion blur, and left-right motion blur-at 5 severity levels to the same validation split, yielding 104,800 corrupted images in total. GTSRB-C is used solely for robustness evaluation.

Table 1. Summary of datasets used in this work.

Dataset	classes	Train	Val	Test	Image Size
GTSRB [18]	43	31,367	7,842	12630	32×32
GTSRB-C	43	—	—	104,800	32×32

4.2 Experimental Details

We describe the network architecture, implementation details, and corruption-aware training strategy below.

Network Architecture. The proposed model adopts a lightweight residual backbone comprising a stem block followed by three residual stages with 64, 128, and 256 output channels, respectively. Each residual block integrates the proposed MCSCA module. A multi-scale feature aggregation head concatenates global average-pooled representations from the second and third stages, yielding a 384 dimensional feature vector. The classifier consists of a BatchNorm layer, a Dropout layer (rate 0.4), and a fully connected layer with 43 output units. An SEBlock is embedded in the stem for early channel recalibration.

Implementation Details. All models are trained for 60 epochs on a single NVIDIA GPU with mixed-precision (AMP) enabled. The batch size is 64 for training and 256 for evaluation, with 4 parallel data-loading workers. The optimizer is AdamW with an initial learning rate of 10^{-3} and weight decay of 2×10^{-4}, scheduled by warmup-cosine annealing with a 5-epoch linear warmup followed by cosine decay. Gradient clipping with maximum norm 5.0 is applied throughout. CutMix and Mixup [25,27] are applied with equal probability during the first 40 epochs using $\lambda \sim \mathrm{Beta}(0.3,0.3)$. Cross-entropy loss with label smoothing $\varepsilon = 0.1$ is adopted. To ensure a fair comparison, all methods share the same backbone, training pipeline, data split (80/20 with fixed random seed), and hyperparameter configuration. The single-branch variants differ

from the complete MCSCA only in the attention module embedded within each residual block.

Table 2. Recognition performance on the GTSRB validation set under clean conditions.

Method	Attention	Acc(%)	Prec(%)	Rec(%)	Params(M)	FPS
Baseline	None	99.89	99.96	99.74	4.34	910.5
NSE [24]	NSE	**99.92**	**99.97**	**99.87**	4.34	660.5
SCC [15]	SCC	**99.92**	**99.97**	**99.87**	4.38	646.7
CDI [11]	CDI	99.89	99.93	99.78	4.35	465.6
MCSCA (Ours)	—	99.89	**99.97**	99.80	4.38	633.4

4.3 Results Analysis

Comparison with Single-branch Variants. As shown in Table 2 and illustrated in Fig. 3, all attention equipped configurations achieve competitive recognition performance relative to the non-attention baseline (99.89% accuracy), confirming the effectiveness of attention-based feature refinement within the proposed backbone. Among the single-branch variants, NSE and SCC each achieve the highest accuracy of 99.92%, while the CDI branch attains 99.89%, matching the baseline. These results indicate that individual attention branches can effectively enhance discriminative feature representations under clean visual conditions when trained with sufficient augmentation.

Figure 4 shows the validation accuracy curves over 60 training epochs. All attention equipped models converge faster than the baseline in early epochs, with MCSCA achieving the most rapid initial improvement, demonstrating the benefit of multi-branch collaborative feature refinement during training.

Analysis of Multi-branch Collaborative Fusion. MCSCA achieves 99.89% accuracy with 99.97% precision and 99.80% recall. Although accuracy is comparable to single-branch variants, the full metric profile shows a more balanced precision – recall trade-off, indicating that adaptive fusion effectively integrates complementary branch behaviors.

Analysis of learned fusion weights reveals clear depth-dependent specialization: CDI contributes more in deeper layers, NSE is more important at intermediate stages, while SCC remains relatively stable across the network. This demonstrates that different attention branches are automatically allocated to different representation levels during training.

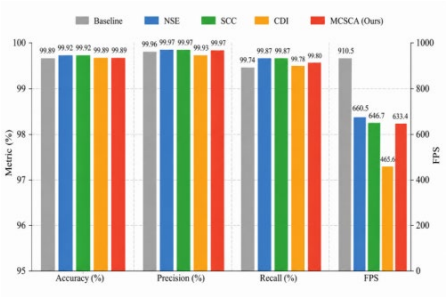

Fig. 3. Performance comparison of different attention strategies on GTSRB. Accuracy, Precision, and Recall are reported on the left y-axis; FPS is reported on the right y-axis.

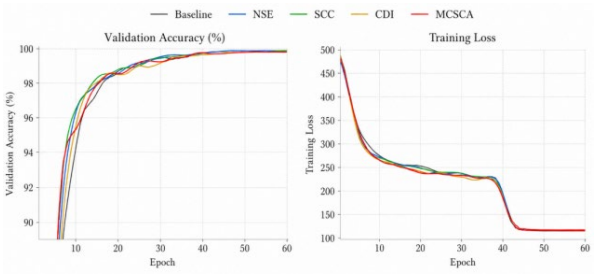

Fig. 4. Validation accuracy curves on GTSRB over 60 training epochs for all five configurations. MCSCA achieves the fastest convergence in early epochs.

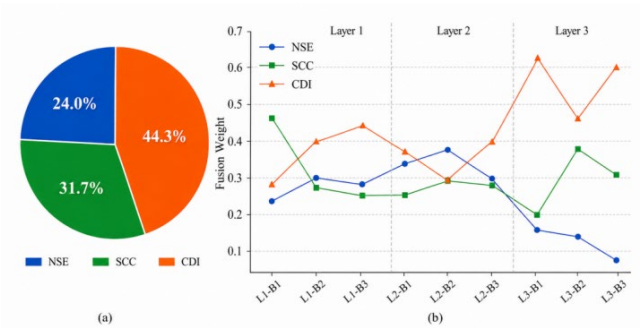

Fig. 5. Branch fusion weight analysis of MCSCA. (a) Mean adaptive Softmax weights averaged across all 9 residual blocks; CDI receives the highest mean weight (44.3%). (b) Per-block branch weight distribution across all 9 residual blocks. CDI weight rises markedly in Layer 3 while NSE weight declines, reflecting depth-dependent specialisa- tion learned end-to-end.

Furthermore, the parameter count of MCSCA (4.38M) is comparable to the single-branch variants (4.34-4.38M), demonstrating that the three-branch collaborative design does not incur substantial additional parameter overhead.

Computational Efficiency. The non-attention baseline achieves the highest inference speed of 910.5 FPS. Among attention-based variants, NSE (660.5 FPS) and SCC

(646.7 FPS) show moderate computational overhead, while CDI (465.6 FPS) is relatively slower due to additional tensor operations. The complete MCSCA model achieves 633.4 FPS, remaining comparable to single-branch variants while integrating three parallel attention pathways.

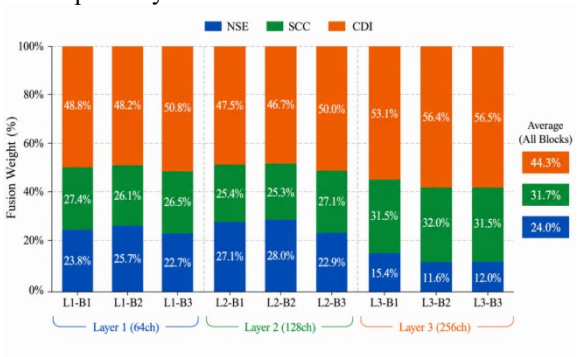

Fig. 6. Per-block branch weight distribution (stacked) across all 9 residual blocks. Dashed lines separate the three residual stages. CDI dominance increases visibly from Layer 1 to Layer 3.

4.4 Robustness Experiments

Evaluation Protocol. Robustness is evaluated on GTSRB-C, constructed by applying 8 camera corruption types at 5 severity levels to the GTSRB validation split, yielding 104,800 corrupted images. Three complementary metrics are reported. (i) *Mean Corruption Accuracy* (mCA, %): average classification accuracy across all five severity levels. (ii) *Relative Performance Change* (rPC): the normalised drop from severity-1 to severity-5 accuracy, $\mathrm{rPC} = \max(0, (a_1 - a_5)/a_1)$, where $\mathrm{rPC} = 0$ indicates no degradation. (iii) *Composite Robustness Score* (RobScore):

$$\mathrm{RobScore} = 0.75 \times \mathrm{mCA} + 0.25 \times (1 - \mathrm{rPC}) \times 100 \tag{29}$$

which jointly penalises low mean accuracy and high severity sensitivity.

Results. Table 3 reports per-corruption performance. Fig. 7 shows per-severity accuracy, and Fig. 8 summarizes overall robustness across corruption types, including photometric, noise, and motion/occlusion degradations.

Photometric Corruptions. MCSCA achieves near-perfect robustness under photometric distortions. Fog reaches 99.85% mCA with rPC = 0, and rain achieves 98.87% mCA with rPC = 0, indicating stable performance across all severity levels. This robustness is largely attributed to strong color jitter and random erasing during training, which implicitly simulate illumination shifts.

Additive Noise Corruptions. The model maintains high performance at low severity but degrades under strong noise, with impulse noise being the most challenging due to severe distortion of fine-grained features.

Overall Robustness. Across all corruptions, the model achieves 81.81% and mean RobScore of 76.91. Performance follows the trend: photometric corruptions $\gg$ motion corruptions $\approx$ additive noise $\gg$ snow, consistent with the training augmentations. While photometric and mild motion distortions are well covered, severe occlusion remains an open challenge.

Table 3. Camera corruption robustness on GTSRB-C.

Category	Corruption	mCA(%)	rPC	RobScore
Photometric	Fog	99.85	0.000	99.89
	Rain	98.87	0.000	99.15
Additive Noise	Uniform	79.40	0.377	75.13
	Gaussian	77.71	0.452	71.99
	Impulse	71.05	0.662	61.74
Motion / Occlusion	Motion (FB)	81.83	0.427	75.69
	Motion (LR)	80.68	0.423	74.94
	Snow	65.08	0.682	56.76
Mean	—	**81.81**	**0.328**	**76.91**

Fig. 7. Per-severity accuracy (%) heatmap on GTSRB-C across 8 corruption types and 5 severity levels. Green indicates high accuracy; red indicates severe degradation.

5 Conclusion

In this paper, we propose MCSCA, a Multi-dimensional Collaborative Spatial-Channel Attention network for robust traffic sign recognition, which integrates three complementary attention branches-NSE, SCC, and CDI-in parallel through a learnable Softmax-normalized adaptive fusion strategy with residual connection. Experiments on GTSRB demonstrate 99.89% validation accuracy, 99.97% precision, and 99.80% recall at 633.4 FPS with 4.38M parameters. Analysis of the learned fusion weights reveals depth-dependent specialisation: CDI weight increases from 0.372 in Layer 1 to 0.563 in Layer 3, while NSE peaks at intermediate depths and SCC maintains a stable contribution throughout, confirming the complementary nature of the three branches. Robustness evaluation on GTSRB-C yields a mean mCA of 81.81% and RobScore of 76.91, with near-perfect performance under photometric corruptions (Fog: 99.85%, Rain: 98.87%) and graceful degradation under noise and motion blur; snow occlusion remains the primary open challenge. In future work, we plan to incorporate targeted corruption-

aware training and extend evaluation to real-world datasets collected under adverse weather conditions.

References

1. An, F., Wang, J., Liu, R.: Road traffic sign recognition algorithm based on cascade attention-modulation fusion mechanism. IEEE Transactions on Intelligent Trans- portation Systems 25(11), 17841-17851 (2024)
2. Ciregan, D., Meier, U., Schmidhuber, J.: Multi-column deep neural networks for image classification. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3642-3649. IEEE (2012)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 886-893 (2005)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Escalera, S., Pujol, O., Radeva, P.: Traffic sign recognition system with boosted hog features. Machine Vision and Applications 22(1), 99-113 (2011)
6. Farzipour, A., Manzari, O.N., Shokouhi, S.B.: Traffic sign recognition using local vision transformer. In: 2023 13th international conference on computer and knowledge engineering (ICCKE). pp. 191-196. IEEE (2023)
7. Greenhalgh, J., Mirmehdi, M.: Real-time detection and recognition of road traffic signs. IEEE Transactions on Intelligent Transportation Systems 13(4), 1498-1506 (2012)
8. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7132-7141 (2018)
9. Maldonado-Bascón, S., Lafuente-Arroyo, S., Gil-Jimenez, P., Gómez-Moreno, H., López-Ferreras, F.: Road-sign detection and recognition based on support vector machines. IEEE transactions on intelligent transportation systems 8(2), 264 – 278 (2007)
10. Min, W., Liu, R., He, D., Han, Q., Wei, Q., Wang, Q.: Traffic sign recognition based on semantic scene understanding and structural traffic sign location. IEEE Transactions on Intelligent Transportation Systems 23(9), 15794-15807 (2022)
11. Misra, D., Nalamada, T., Arasanipalai, A.U., Hou, Q.: Rotate to attend: Convolutional triplet attention module. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 3139-3148 (2021)
12. Naz, I., Shah, J.H., Tahir, A., Marri, M.R., Saleem, R., Saeed, M.E.S.: Attention to detail: A conditional multi-head transformer for traffic sign recognition. Plos one 20(11), e0335341 (2025)
13. Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., Ye, Q.: Conformer: Local features coupling global representations for visual recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 367-376 (2021)
14. Sermanet, P., LeCun, Y.: Traffic sign recognition with multi-scale convolutional networks. In: The 2011 international joint conference on neural networks. pp. 2809-2813. IEEE (2011)
15. Si, Y., Xu, H., Zhu, X., Zhang, W., Dong, Y., Chen, Y., Li, H.: Scsa: Exploring the synergistic efects between spatial and channel attention. Neurocomputing 634, 129866 (2025)
16. Song, S., Ye, X., Manoharan, S.: E-mobilevit: a lightweight model for traffic sign recognition. Industrial Artificial Intelligence 3(1), 3 (2025)

17. Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C.: The german traffic sign recognition benchmark: a multi-class classification competition. In: The 2011 international joint conference on neural networks. pp. 1453-1460. IEEE (2011)
18. Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C.: Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural networks 32, 323-332 (2012)
19. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137-154 (2004)
20. Wang, F., Bai, J., Wang, M., Liu, B., Xue, H., Chen, J.: Robust traffic sign detection in real-world harsh conditions: A pioneering benchmark dataset and attention-based methodology. Engineering Applications of Artificial Intelligence 166, 113526 (2026)
21. Wang, H.: Traffic sign recognition with vision transformers. In: Proceedings of the 6th International Conference on Information System and Data Mining. pp. 55-61 (2022)
22. Wang, T., Hu, Y., Fang, Q., He, B., Gong, X., Wang, P.: Dk-former: A hybrid structure of deep kernel gaussian process transformer network for enhanced traffic sign recognition. IEEE Transactions on Intelligent Transportation Systems 25(11), 18561-18572 (2024)
23. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3-19 (2018)
24. Yang, L., Zhang, R.Y., Li, L., Xie, X.: Simam: A simple, parameter-free attention module for convolutional neural networks. In: International conference on machine learning. pp. 11863 – 11874. PMLR (2021)
25. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Training strategy that makes use of sample pairing. In: ICCV (2019)
26. Zaklouta, F., Stanciulescu, B.: Real-time traffic sign recognition in three stages. Robotics and Autonomous Systems 62(1), 16-24 (2014)
27. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. ICLR (2018)
28. Zhang, Y., Lu, Y., Zhu, W., Wei, X., Wei, Z.: Traffic sign detection based on multiscale feature extraction and cascade feature fusion: Y. zhang et al. The Journal of Supercomputing 79(2), 2137-2152 (2023)
29. Zheng, Y., Jiang, W.: Evaluation of vision transformers for traffic sign classification. Wireless Communications and Mobile Computing 2022(1), 3041117 (2022)
30. Zhu, Z., Liang, D., Zhang, S., Huang, X., Li, B., Hu, S.: Traffic-sign detection and classification in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2110-2118 (2016)
31. Li, Z., Hu, Y., Chen, Z., Huang, Q., Qiu, G., Fu, Z., & Liu, M. (2026, March). Retrack: Evidence-driven dual-stream directional anchor calibration network for composed video retrieval. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 28, pp. 23373-23381).
32. Li, Z., Hu, Y., Chen, Z., Zhang, M., Fu, Z., & Nie, L. (2026). Conesep: Cone-based robust noise-unlearning compositional network for composed image retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16897-16909).
33. Zhang, J., & Lu, G. (2024). Underground mapping and localization based on ground-penetrating radar. In Proceedings of the Asian Conference on Computer Vision (pp. 2018-2033).