CCSPNet-Joint: Efficient Joint Training Method for Traffic Sign Detection Under Extreme Conditions

1st Haoqin Hong
Institute of Innovation & Entrepreneurship, Hanhong College
Southwest University
Chongqing, China
honghaoqin@email.swu.edu.cn

2nd Yue Zhou (✉)
School of Artificial Intelligence
Southwest University
Chongqing, China
zhouyuenju@163.com

3rd Xiangyu Shu
College of Computer and Information Science
Southwest University
Chongqing, China
shuxy6263@163.com

4th Xiaofang Hu
School of Artificial Intelligence
Southwest University
Chongqing, China
huxf@swu.edu.cn

2024 International Joint Conference on Neural Networks (IJCNN) | 979-8-3503-5931-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCNN60899.2024.10651346

***Abstract*—Traffic sign detection is an important research direction in intelligent driving. Unfortunately, existing methods often overlook training methods specifically designed for extreme conditions such as fog, rain, and motion blur. Moreover, the end-to-end training strategy for image denoising and object detection models fails to utilize inter-model information effectively. To address these issues, we propose CCSPNet, an efficient feature extraction module based on Contextual Transformer and CNN, capable of effectively utilizing the static and dynamic features of images, achieving faster inference speed and providing stronger feature enhancement capabilities. Furthermore, we establish the correlation between object detection and image denoising tasks and propose a joint training model, CCSPNet-Joint, to improve data efficiency and generalization. Finally, to validate our approach, we create the CCTSDB-AUG dataset for traffic sign detection in extreme scenarios. Extensive experiments have shown that CCSPNet achieves state-of-the-art performance in traffic sign detection under extreme conditions. Compared to end-to-end methods, CCSPNet-Joint achieves a 5.32% improvement in precision and an 18.09% improvement in mAP@.5.**

***Index Terms*—Traffic sign detection, Joint training method**

I. Introduction

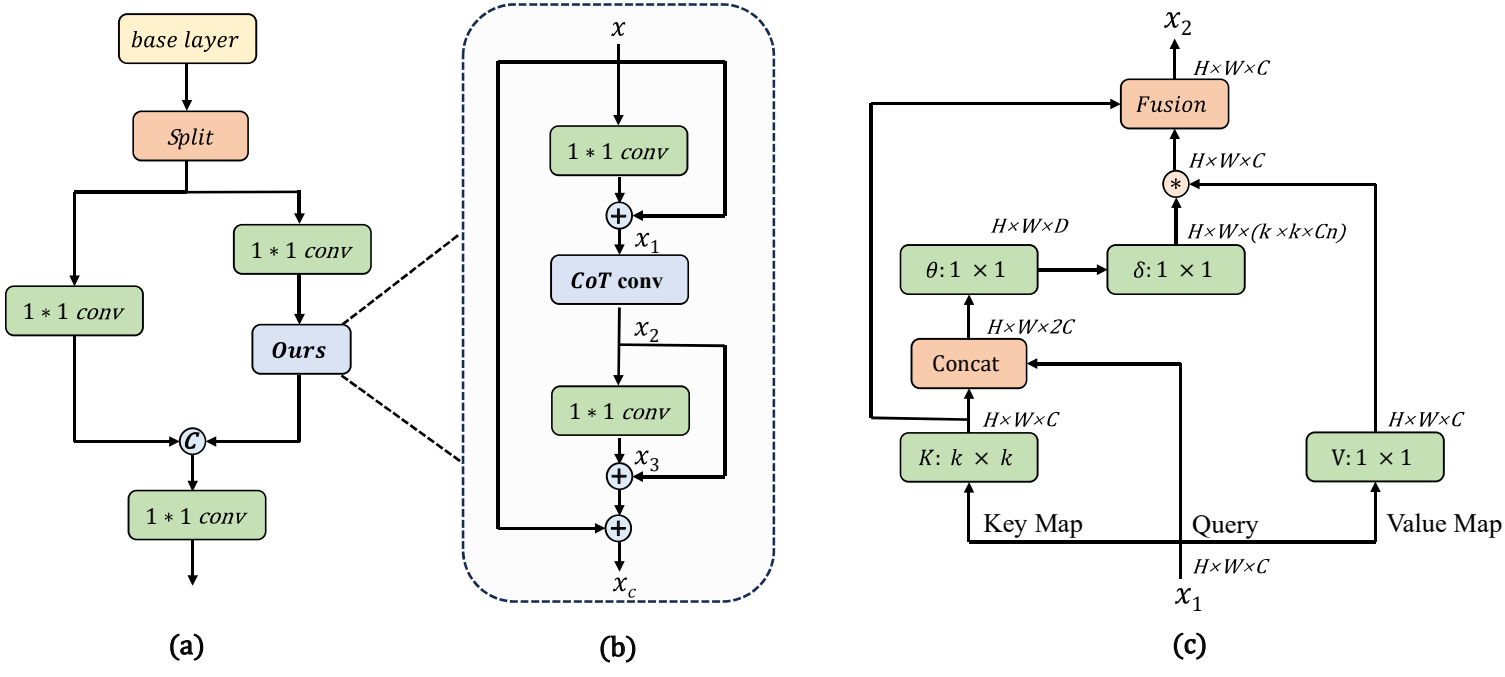
Traffic sign detection (TSD) plays a significant role in the field of intelligent driving by providing vital road information to intelligent driving systems, enabling accurate recognition for subsequent decision-making processes[7, 32]. Traffic sign detection algorithms utilize computer vision techniques to rapidly and accurately identify and extract information from images or video data pertaining to traffic signs[24]. The application of such algorithms aids intelligent vehicles in the real-time acquisition of road sign information, enhancing driving safety and overall driving efficiency.

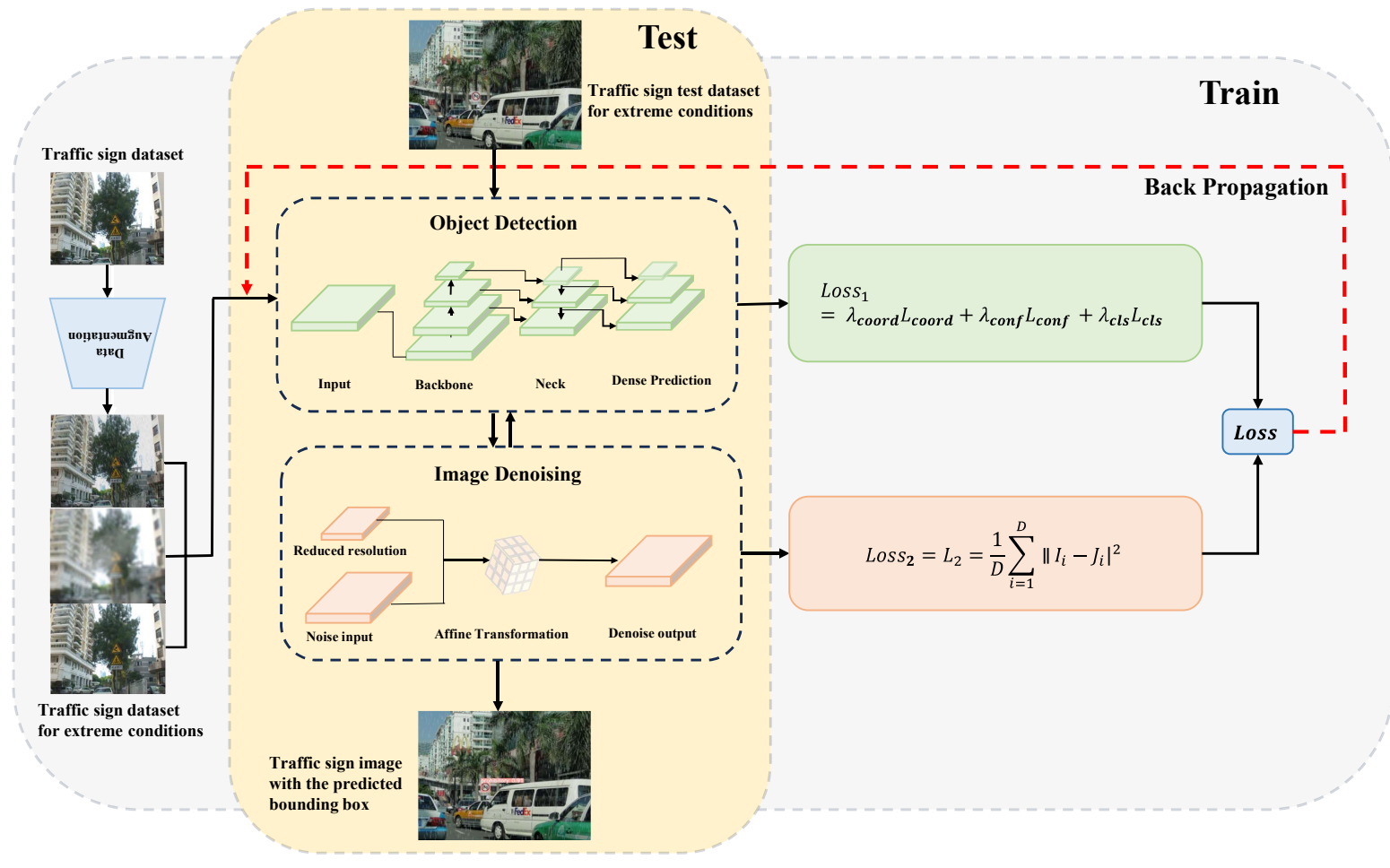
Early traffic sign detection algorithms primarily relied on traditional computer vision techniques, such as image edge detection [21], image filtering algorithms [10], and morphological image processing [6]. With the advancement of deep learning technology[15], deep learning-based traffic sign detection algorithms have gained widespread adoption. Applying deep learning methods to autonomous driving traffic sign detection improves detection accuracy, speed, and enhances the model's generalization capability. Convolutional neural networks (CNNs) demonstrate strong advantages in traffic sign detection. Compared to other algorithms, CNNs extract more comprehensive features, effectively improving recognition accuracy and speed. CNNs process images through convolution and pooling operations, extracting different features, and automatically learning optimal features, thereby enhancing the accuracy of traffic sign detection. Additionally, CNNs exhibit adaptability, allowing for adjustments in network depth and structure to accommodate different traffic sign detection tasks[14]. However, CNN-based methods alone may not effectively capture global information in complex traffic scenarios. Therefore, it is crucial to design an effective model that can efficiently extract local feature information while improving the capture of global information in traffic scenes.

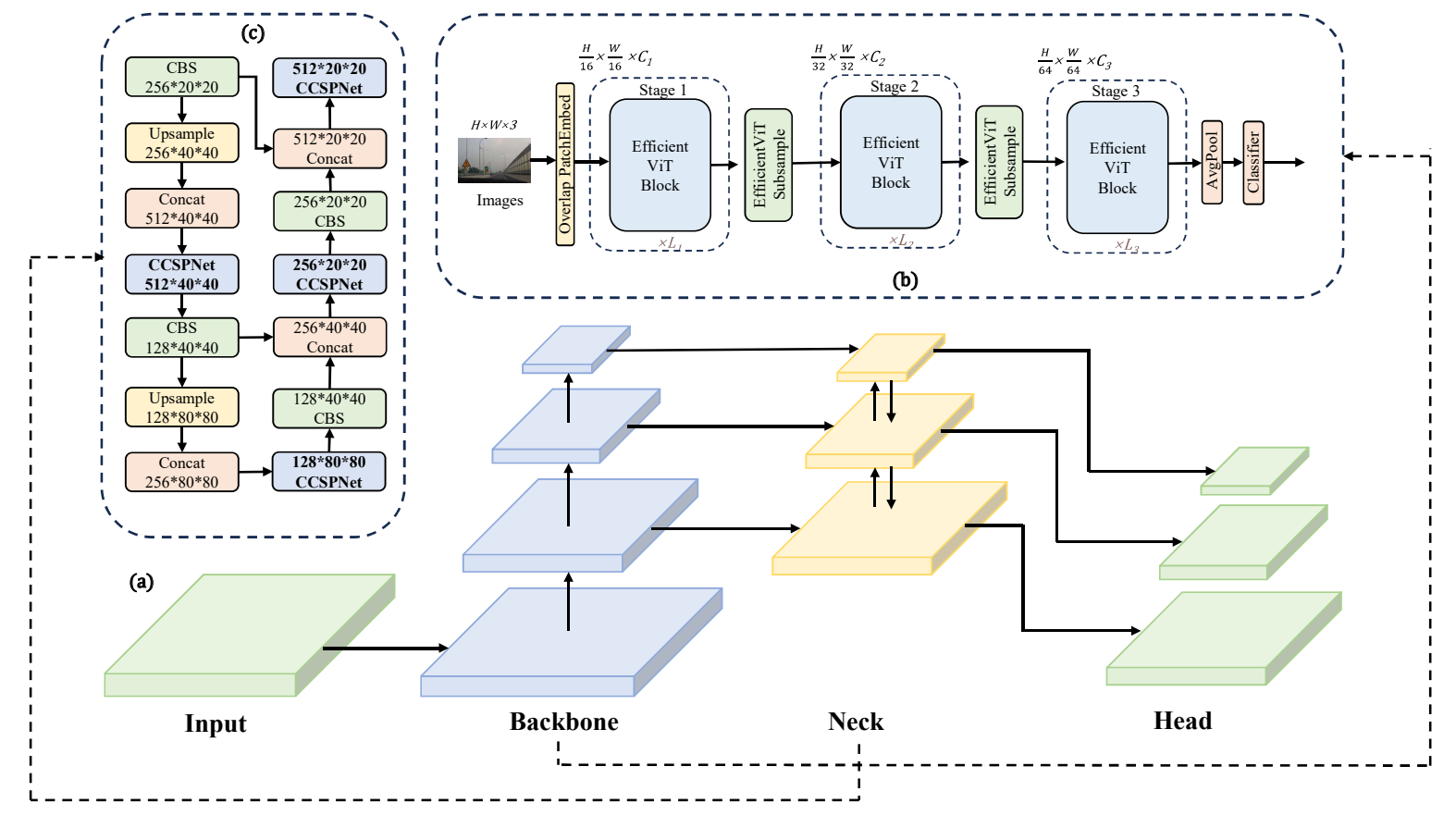
The backbone network based on Transformer [27] has better capabilities in capturing global information compared to the backbone network based on CNNs. Transformer has been widely used in various visual tasks, including image classification [5, 30], object detection [2], and semantic segmentation[19, 18], among others, for feature extraction. The advantage of Transformer lies in its ability to handle long-range dependencies and parallel computation, effectively capturing crucial information in input sequences. It enables end-to-end training and inference for different tasks. However, existing traffic sign detection models have not fully utilized the advantages of Transformer. In practical traffic scenarios, extracting global features allows for comprehensive consideration of overall shape, color, and other information related to traffic signs. These features possess certain scale and rotation

Our code is at https://github.com/HaoqinHong/CCSPNet-Joint.
The authors 2nd and 4th are also with Chongqing Key Laboratory of Brain-inspired Computing and Intelligent Chips.

Authorized licensed use limited to: Southwest University. Downloaded on September 15,2024 at 08:06:20 UTC from IEEE Xplore. Restrictions apply.


base layer
Split
1 * 1 conv
1 * 1 conv
Ours
C
1 * 1 conv
x
1 * 1 conv
x1
CoT conv
x2
1 * 1 conv
x3
xc
(a)
(b)
x2
H×W×C
Fusion
H×W×C
*
H×W×D
θ: 1 × 1
H×W×(k ×k×Cn)
δ: 1 × 1
H×W×2C
Concat
H×W×C
K: k × k
H×W×C
V: 1 × 1
Key Map
Query
Value Map
H×W×C
x1
(c)


Test
Traffic sign test dataset for extreme conditions
Train
Traffic sign dataset
Data Augmentation
Traffic sign dataset for extreme conditions
Back Propagation
Object Detection
Input
Backbone
Neck
Dense Prediction
$Loss_1 = \lambda_{coord}L_{coord} + \lambda_{conf}L_{conf} + \lambda_{cls}L_{cls}$
Loss
Image Denoising
Reduced resolution
Noise input
Affine Transformation
Denoise output
$Loss_2 = L_2 = \frac{1}{D}\sum_{i=1}^{D} \| I_i - J_i \|^2$
Traffic sign image with the predicted bounding box


(c)
CBS 256*20*20
512*20*20 CCSPNet
Upsample 256*40*40
512*20*20 Concat
Concat 512*40*40
256*20*20 CBS
CCSPNet 512*40*40
256*20*20 CCSPNet
CBS 128*40*40
256*40*40 Concat
Upsample 128*80*80
128*40*40 CBS
Concat 256*80*80
128*80*80 CCSPNet
H×W×3
Images
Overlap PatchEmbed
Stage 1
Efficient ViT Block
×L1
EfficientViT Subsample
Stage 2
Efficient ViT Block
×L2
EfficientViT Subsample
Stage 3
Efficient ViT Block
×L3
AvgPool
Classifier
(b)
(a)
Input
Backbone
Neck
Head