ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Research on Deep Learning-Based Methods for Matching Traffic Sign Images with Textual Captions

Ning Wang[1] and Jian Qu[2]

ABSTRACT

Intelligent transportation systems face challenges in matching traffic sign images with natural language descriptions, particularly in modal heterogeneity and fine-grained semantic alignment. This issue is crucial for accurately understanding the traffic environment and safety decision-making in autonomous driving, carrying significant application value. Most existing methods are based on classification or template matching and lack deep semantic modelling between images and texts, making it difficult to adapt to real-world complex scenarios. To address this, this paper proposes a deep learning-based image-text matching method that automatically extracts directory structures to generate fine-grained labels, and introduces InfoNCE contrastive loss based on intra-batch negative samples to achieve cross-modal learning. Pré-trained ResNeXt50_32x4d and DistilBERT are employed as image and text encoders, which are uniformly mapped to a shared embedding space. Experimental results demonstrate that the proposed method outperforms existing methods regarding Recall@1, mean Average Precision (mAP), and Mean Reciprocal Rank (MRR), showcasing stronger semantic alignment capability and application potential.

Article information:

Keywords: Traffic Sign Recognition, Cross-modal Matching, Contrastive Learning, InfoNCE Loss

Article history:
Received: May 15, 2025
Revised: July 7, 2025
Accepted: July 7, 2025
Published: July 19, 2025
(Online)

DOI: 10.37936/ecti-cit.2025193.262155

1. INTRODUCTION

1.1 Research Background

Traffic signs are key information in road traffic management and autonomous driving systems, and a correct knowledge of traffic signs can effectively ensure road safety and intelligent decision-making. With the rapid development of autonomous driving technology, accurate recognition and understanding of traffic signs are becoming increasingly important [1]. However, in complex traffic scenarios, existing studies only identify the categories of signs, and the text description of traffic conditions accounts for the environment. Still, there is no specific description of traffic signs, which means it does not meet actual needs, so we need to understand the particular meaning and indication content of the signs. On the night of 29 March 2025, a traffic accident caused social concern, demonstrating the importance of this issue. Three university students drove a Xiaomi SU7 from Hubei to Anhui Province and suffered a serious traffic accident on the Zongyang-Qimen section of the Deshang Expressway. At that time, the vehicle was in NOA intelligent driving assistance mode, and the speed was 116 kilometres per hour. Due to road construction, a guardrail blocked the lane, forcing cars the oncoming lane. According to the news report, the system issued an early warning after detecting an obstacle. It began to slow down, the driver immediately took over the vehicle, switched to manual mode, further slowed down, adjusted the direction. However, due to the issue of takeover time and the lack of understanding of the road conditions, the vehicle still collided with the concrete column on the guardrail. The accident illustrates the urgent need for interpretable road signs and traffic status descriptions in autonomous driving. Traditional traffic sign recognition methods mainly focus on single-modal image classification. Still, in practical applications, combining information such as text interpretation of signs is often necessary to assist drivers in the real-time understanding of road conditions and reasonable, intel-

[1,2] The authors are with The Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi, Thailand, 11120, Email: 6372100142@stu.pim.ac.th and jianqu@pim.ac.th

[2] Corresponding author: jianqu@pim.ac.th

ligent decision-making. Cross-modal matching aims to map images and text to a shared feature space for efficient retrieval and matching [2]. Although the progress of contrastive learning and dual-stream network architecture has promoted the development of this field, how to effectively use fine-grained category information in complex traffic scenarios is still an urgent problem to be solved [3].

1.2 Research Motivation and Issues

Currently, most recognition and text-matching methods often have category labels with manual annotation, leading to high annotation costs and limited category granularity [4]. However, subtle differences between traffic signs can significantly impact driving decisions, so detailed classification and description are significant. In this study, the data were first manually classified. Then, fine-grained category labels were automatically generated by extracting the hierarchical structure information of traffic sign folders to reduce the cost of labelling and improve the meticulousness and accuracy of categories. At the same time, in cross-modal matching, how to effectively construct negative sample pairs and address the issue of overly simple positive sample matching in the same batch is another critical challenge [5]. Traditional methods often rely on hand-designed negative sample selection strategies or large amounts of pre-labeled data. To solve this problem, an InfoNCE loss function based on intra-batch negative samples was introduced, using all unmatched sample pairs in the same batch as negative samples. This method enhances the model's discriminative ability and generalisation performance [6].

1.3 Contributions

The main contributions of this study are:

We propose a method to automatically generate fine-grained category labels by parsing the directory structure of datasets. This method can extract image categories and associate them with the corresponding text descriptions, improving the category granularity and effectively reducing manual annotation costs.

After collecting image data from our model car, we created a new set of multimodal traffic sign datasets with 15,200 images and 30,400 text descriptions. For each image, we generated two sets of text descriptions.

We designed a cross-modal matching model based on contrastive learning. The model uses pre-trained Resnext50_32x4d as the image encoder and DistilBERT as the text encoder.It introduces the InfoNCE loss based on intra-batch negative samples, effectively improves its discrimination ability.

Experiments are carried out on the constructed traffic sign dataset, and the proposed method is superior to other existing model combinations in terms of Recall@K, MRR, and mAP. In addition, the model's internal mechanism and performance characteristics were further demonstrated through visual analysis.

2. RELATED WORK

2.1 Improved Traffic Sign Recognition and Image-Text Matching

In recent years, traffic sign recognition has made significant progress. Liang et al. achieved high-precision sign detection based on improved YOLOv5, but their method was limited to image recognition and did not involve alignment and matching with text semantics [7]. In addition, existing research also has deficiencies at the data level. Mainstream data sets like GTSRB and TT100K have limited coverage categories and lack multimodal semantic annotations [8]. In addition, image description technology has been used to describe the collected road scenes in natural language to assist the system in understanding the environment. However, existing research focuses on driving behavior and the overall scene, lacking a dedicated description of details such as road conditions, obstacles, and traffic signs [9]. For example, Wei Li *et al.* proposed a traffic scene understanding method based on image description in 2020, generating driving suggestions and strategies through image description and using image description technology to develop a natural language description of traffic scenes to assist advanced driver assistance systems (ADAS) in making decisions under special traffic conditions [10]. However, this experiment mainly focuses on the description of driving behavior and the overall scene. It lacks a detailed description of specific road conditions and traffic signs, so there are still deficiencies in the dedicated description of specific road conditions, obstacles, and traffic signs in autonomous driving [11].

To this end, this paper constructs a new, comprehensive, multi-angle feature-based image-text matching traffic sign dataset that specifically targets road conditions and the presence of obstacles and focuses on traffic signs. It also proposes an image-text matching model that combines ResNeXt50_32x4d and DistilBERT and introduces InfoNCE contrastive learning loss to solve these problems [12].

2.2 Fine-Grained Understanding and Efficient Annotation

In terms of fine-grained semantic understanding, traditional methods, such as the GTSRB dataset, only rely on simple category labels and do not have the ability to delve deeply into the attributes of the traffic sign (such as location, shape, colour, etc.), resulting in the limited capacity of the model to understand complex scenes [13]. Regarding annotation efficiency, traditional methods often rely on manual annotation of the marks in the images one by one, espe-

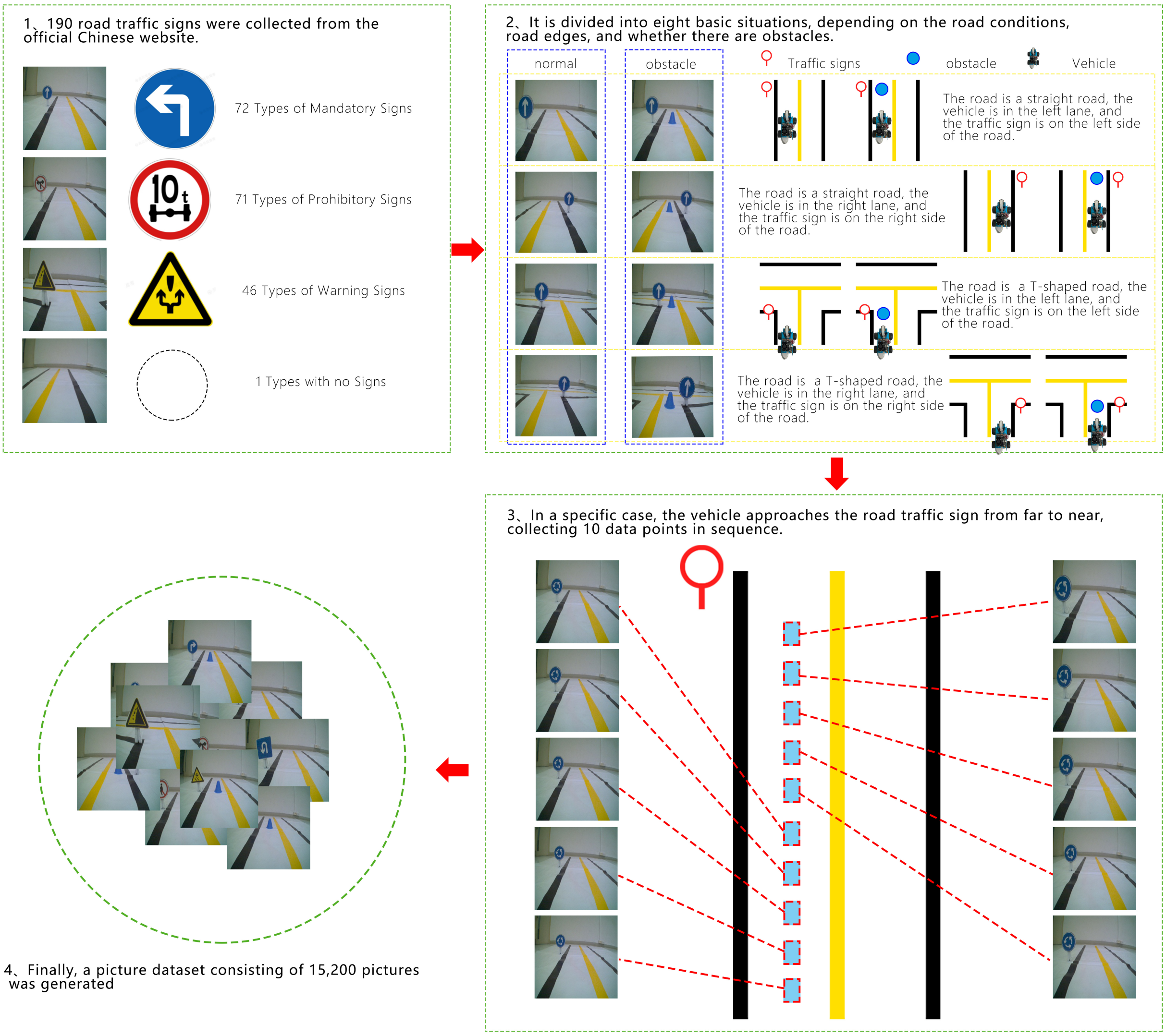
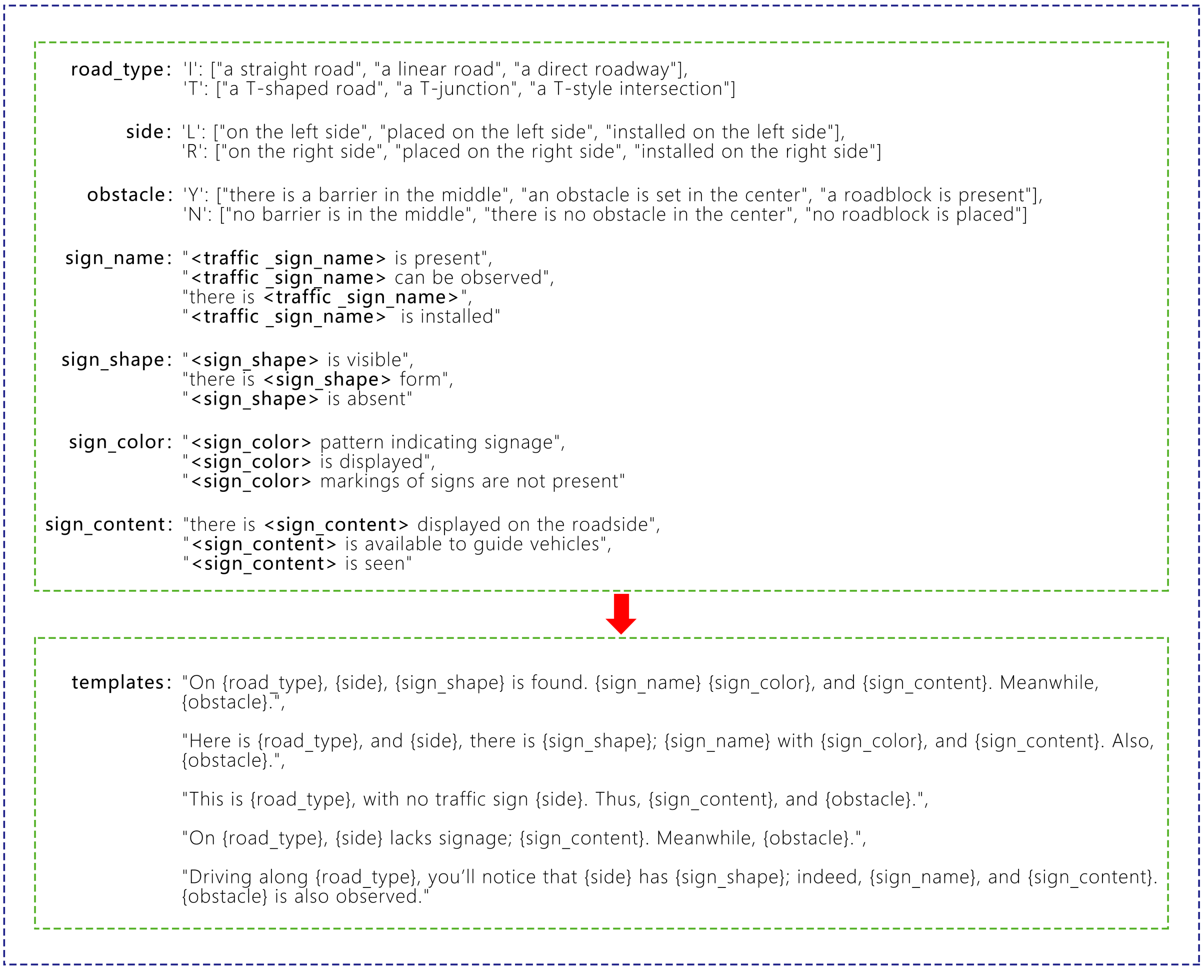

1、190 road traffic signs were collected from the official Chinese website.
72 Types of Mandatory Signs
71 Types of Prohibitory Signs
46 Types of Warning Signs
1 Types with no Signs
2、It is divided into eight basic situations, depending on the road conditions, road edges, and whether there are obstacles.
normal
obstacle
Traffic signs
obstacle
Vehicle
The road is a straight road, the vehicle is in the left lane, and the traffic sign is on the left side of the road.
The road is a straight road, the vehicle is in the right lane, and the traffic sign is on the right side of the road.
The road is a T-shaped road, the vehicle is in the left lane, and the traffic sign is on the left side of the road.
The road is a T-shaped road, the vehicle is in the right lane, and the traffic sign is on the right side of the road.
3、In a specific case, the vehicle approaches the road traffic sign from far to near, collecting 10 data points in sequence.
4、Finally, a picture dataset consisting of 15,200 pictures was generated

Fig.1: *Image data collection flow chart.*

cially in the construction of large-scale datasets; the annotation process is cumbersome, time-consuming, and prone to errors. For example, the TT100K dataset relies on manual labelling of the category and location of each mark, and the labelling process lacks standardisation, which can lead to inaccuracies and inconsistencies. In addition to the manual annotation of each piece of data, it is also necessary to collect and organise it, and even a complete check is required to prevent errors. Therefore, the traditional traffic sign recognition methods have shortcomings in fine-grained semantic understanding and labelling efficiency [14]. To address these issues, this paper constructs a hierarchical label system to accurately describe multiple attributes of signs, improving traffic signs' recognition accuracy in complex scenes. At the same time, an automatic labelling method based on directory structure is proposed, which enhances labelling efficiency, reduces labour costs, and improves the accuracy and consistency of labelling to a certain extent by automatically generating labelling information.

2.3 Contrastive Learning and the InfoNCE Loss

Contrastive learning is a self-supervised learning method that learns discriminative feature representations by pulling pairs of positive samples closer and negative samples farther away. In the cross-modal learning task, contrastive learning provides an effective framework for semantic information alignment between different modalities [15]. The InfoNCE (Noise Comparative Estimation Based on Mutual Information) loss function is a commonly used form of contrastive learning, which was first proposed by van den Oord *et al.* in Contrastive Predictive Coding (CPC). The mathematical expression is as follows:



```
    road_type: 'I': ["a straight road", "a linear road", "a direct roadway"],
               'T': ["a T-shaped road", "a T-junction", "a T-style intersection"]

         side: 'L': ["on the left side", "placed on the left side", "installed on the left side"],
               'R': ["on the right side", "placed on the right side", "installed on the right side"]

     obstacle: 'Y': ["there is a barrier in the middle", "an obstacle is set in the center", "a roadblock is present"],
               'N': ["no barrier is in the middle", "there is no obstacle in the center", "no roadblock is placed"]

    sign_name: "<traffic _sign_name> is present",
               "<traffic _sign_name> can be observed",
               "there is <traffic _sign_name>",
               "<traffic _sign_name>  is installed"

   sign_shape: "<sign_shape> is visible",
               "there is <sign_shape> form",
               "<sign_shape> is absent"

   sign_color: "<sign_color> pattern indicating signage",
               "<sign_color> is displayed",
               "<sign_color> markings of signs are not present"

 sign_content: "there is <sign_content> displayed on the roadside",
               "<sign_content> is available to guide vehicles",
               "<sign_content> is seen"

    templates: "On {road_type}, {side}, {sign_shape} is found. {sign_name} {sign_color}, and {sign_content}. Meanwhile,
               {obstacle}.",

               "Here is {road_type}, and {side}, there is {sign_shape}; {sign_name} with {sign_color}, and {sign_content}. Also,
               {obstacle}.",

               "This is {road_type}, with no traffic sign {side}. Thus, {sign_content}, and {obstacle}.",

               "On {road_type}, {side} lacks signage; {sign_content}. Meanwhile, {obstacle}.",

               "Driving along {road_type}, you'll notice that {side} has {sign_shape}; indeed, {sign_name}, and {sign_content}.
               {obstacle} is also observed."
```

Fig.2: *Text data creation logic flow chart.*

$$\mathrm{L_{InfoNCE}} = -\mathbb{E}_{(x,y)\sim p_{data}}\left[\log \frac{\exp(f(x,y)/\tau)}{\sum_{y'\in y}\exp(f(x,y')/\tau)}\right] \quad (1)$$

Where $f(x, y)$ is the similarity score between pairs of samples (x, y), τ is the temperature parameter that controls the smoothness of the similarity distribution, and Y is the set containing one positive sample and multiple negative samples.

InfoNCE loss usually occurs in image-matching tasks in both image-to-text and text-to-image directions. For a batch of size N, each image and its corresponding text form a positive sample pair and a negative sample pair with the remaining N-1 texts. Similarly, each piece of text and its corresponding image are positive samples, and the rest are negative. This intra-batch negative sample construction strategy utilizes all unmatched image-text pairs in the batch as negative samples, avoiding additional negative sample sampling steps.

The temperature parameter τ plays a key role in training, determining how "sharp" the similarity distribution is – a smaller τ makes the model more focused on complex negative samples, and a larger τ makes the distribution smoother.

Table1

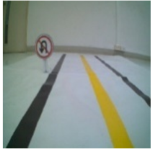
3. MATERIALS AND METHODS

3.1 Datasets and Experimental Setup

3.1.1 Dataset Construction

In this study, we used images of traffic signs captured by the Jetson Nano car. Our dataset consists of 15,200 traffic sign images, each with two different text descriptions, making up a dataset of 30,400 pairs of graphic samples. In the process of data collection, the Jetson Nano car collects data from far and near in two road environments, as well as road signs placed in different positions, and whether there are roadblocks, to ensure that. Image. The data is different. At the same time, the dataset directory structure is designed, which plays an essential role in encoding fine-grained category information. The leading directory is different street sign types, and each leading category directory contains multiple subdirectories that

Table 1: *Details of the data sample with the ID type "055ILN".*

image	captions
	"The 'U-turn prohibited' sign, a round sign and featuring a red border on white, is positioned on the left side of a straight road, signifying that it forbids vehicles from making a U-turn. In addition, there is no roadblock."
	"You observe a round sign, the 'U-turn prohibited' sign featuring a red border on white, located on the left side of a direct roadway; it clarifies that the center shows a U-shaped arrow crossed out by a slash, and no roadblock is placed."
	"You observe a circular sign named 'No U-turn' featuring a red border on white, installed on the left side of a direct roadway, it clarifies that it forbids vehicles from making a U-turn, and no roadblock is placed.",
	"Observing a straight road from installed the left side, one finds a round sign named 'No U-turn' with a red ring and white background that denotes the center shows a U-shaped arrow crossed out by a slash. Plus, there is no obstacle in the center."
	"Observing a straight road from the left side, one finds a round sign, the 'U-turn prohibited' sign with a red ring and white background that denotes the center shows a U-shaped arrow crossed out by a slash. Plus, no barrier is in the middle."
	"On a straight road, placed on the left side, you can see a round sign named 'No U-turn' with a red ring and white background, indicating that the center shows a U-shaped arrow crossed out by a slash. Meanwhile, there is no obstacle in the center."
	"Here is a straight road, and on the left side stands a round sign, the 'U-turn prohibited' sign with a red ring and white background. Notably, it forbids vehicles from making a U-turn, and no roadblock is placed."
	"The 'U-turn prohibited' sign, a round sign with a red ring and white background, is positioned on the left side of a straight road, signifying that it forbids vehicles from making a U-turn. In addition, no barrier is in the middle."
	"A round sign, the 'U-turn prohibited' sign featuring a red border on white; is installed on the left side on a direct roadway. it conveys that vehicles are forbidden from making a U-turn, with no roadblock is placed in place."
	"The sign named 'No U-turn' is a circular sign with a red ring and white background installed on the left side of a direct roadway, meaning the center shows a U-shaped arrow crossed out by a slash. Also, no roadblock is placed."
	"Observing a straight road from installed on the left side, one finds a circular sign, the 'U-turn prohibited' sign featuring a red border on white that denotes it forbids vehicles from making a U-turn. Plus, no roadblock is placed."
	"The 'U-turn prohibited' sign, a round sign featuring a red border on white, is positioned on the left side of a direct roadway, signifying that it forbids vehicles from making a U-turn. In addition, no roadblock is placed."
	"On a straight road, placed on the left side, you can see a circular sign named 'No U-turn' featuring a red border on white, indicating that it forbids vehicles from making a U-turn. Meanwhile, no roadblock is placed."
	"Set on the left side of a linear road is a round sign, the sign named 'No U-turn' with a red ring and white background; it conveys that the center shows a U-shaped arrow crossed out by a slash, with no barrier in the middle."
	"On a direct roadway, on the left side, you can see a circular sign named 'No U-turn' with a red ring and white background, indicating that it forbids vehicles from making a U-turn. Meanwhile, no barrier is in the middle."
	"A round sign named 'No U-turn' featuring a red border on white is found placed on the left side of a straight road, declaring it forbids vehicles from making a U-turn. Moreover, there is no obstacle in the center."
	"On a straight road, installed on the left side, you can see a circular sign named 'No U-turn' with a red ring and white background, indicating that the center shows a U-shaped arrow crossed out by a slash. Meanwhile, there is no obstacle in the center."
	" A round sign, named ' No U-turn ', is installed on the left side of a direct roadway. The sign has a red ring and white background; it conveys that there is a U-shaped arrow crossed out by a slash, with no obstacle in the center. "

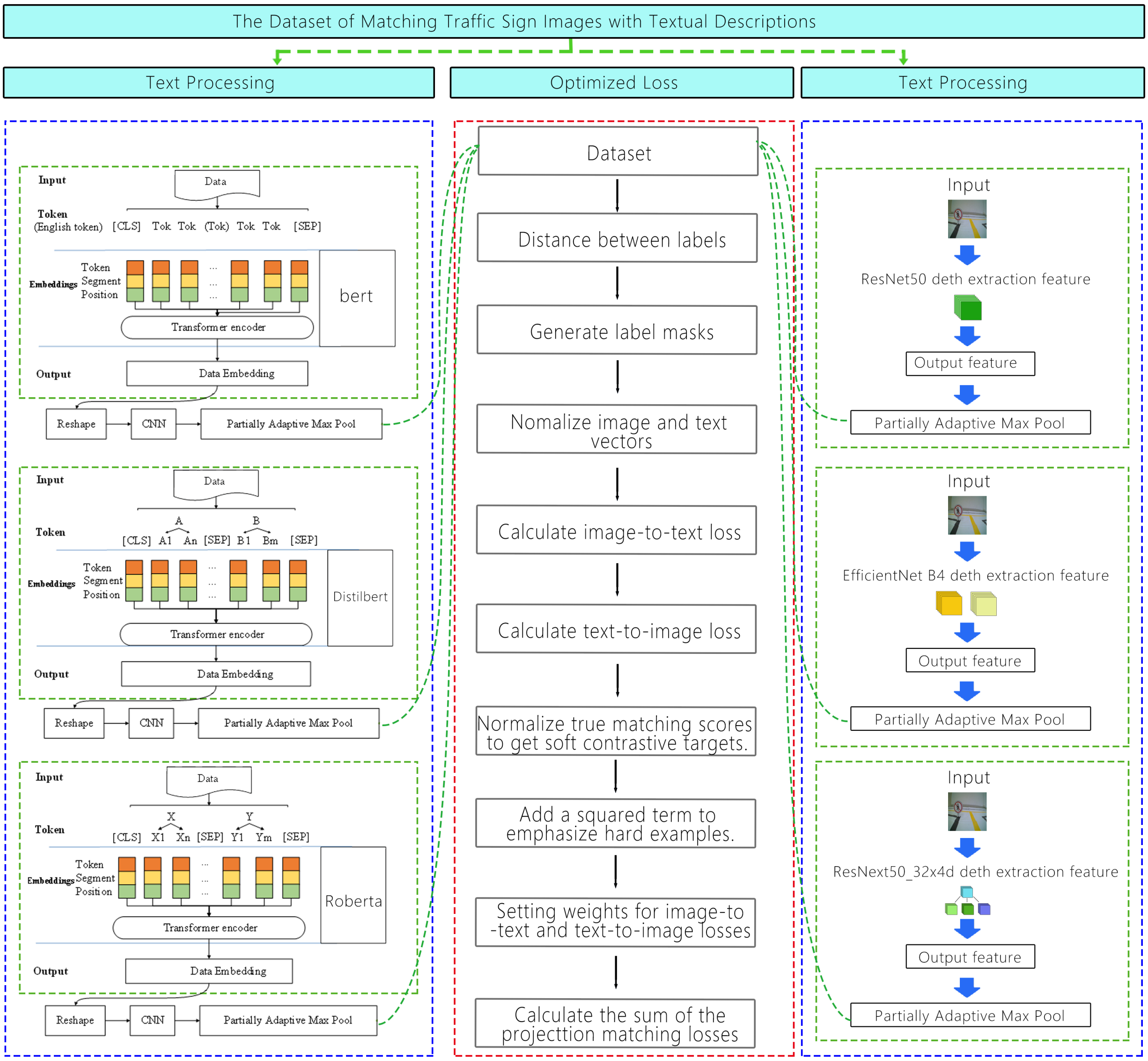

The Dataset of Matching Traffic Sign Images with Textual Descriptions
Text Processing
Optimized Loss
Text Processing
Input
Data
Token
(English token)
[CLS] Tok Tok (Tok) Tok Tok [SEP]
Embeddings
Token
Segment
Position
Transformer encoder
bert
Output
Data Embedding
Reshape
CNN
Partially Adaptive Max Pool
Input
Data
Token
A
B
[CLS] A1 An [SEP] B1 Bm [SEP]
Embeddings
Token
Segment
Position
Transformer encoder
Distilbert
Output
Data Embedding
Reshape
CNN
Partially Adaptive Max Pool
Input
Data
Token
X
Y
[CLS] X1 Xn [SEP] Y1 Ym [SEP]
Embeddings
Token
Segment
Position
Transformer encoder
Roberta
Output
Data Embedding
Reshape
CNN
Partially Adaptive Max Pool
Dataset
Distance between labels
Generate label masks
Nomalize image and text vectors
Calculate image-to-text loss
Calculate text-to-image loss
Normalize true matching scores to get soft contrastive targets.
Add a squared term to emphasize hard examples.
Setting weights for image-to -text and text-to-image losses
Calculate the sum of the projecttion matching losses
Input
ResNet50 deth extraction feature
Output feature
Partially Adaptive Max Pool
Input
EfficientNet B4 deth extraction feature
Output feature
Partially Adaptive Max Pool
Input
ResNext50_32x4d deth extraction feature
Output feature
Partially Adaptive Max Pool

Fig.3: *Traffic sign information retrieval model flow chart.*

identify specific traffic sign attributes. For example, in the case of a straight road and a T-shaped road, the street sign is either on the left or right side of the road, either with a roadblock in front of the road or without a roadblock. This hierarchical structure allows us to extract fine-grained category labels automatically without additional manual annotation. The text description is written according to the traffic laws and standards on the professional platform.

Each image has two different descriptions to enhance linguistic diversity.

The description includes details such as the shape, colour, position, meaning, and application of the traffic signs. Through the interlacing of these details, as well as the reorganisation of the semantics of the language, each text is different, and all text data is stored in JSON format.

3.1.2 Experimental Environment Configuration

The experiments were conducted on a workstation equipped with an NVIDIA A100 GPU and running Linux. The software environment included Python, PyTorch, Transformers, and Timm libraries.

The evaluation metrics are as follows:

- Recall@K (K = 1, 5, 10): the probability of the correct match being found among the top K retrieval results.
- Mean Average Precision (mAP): assesses the overall precision and recall of the retrieval results.
- Mean Reciprocal Rank (MRR): evaluates the rank position of the correct match in the sorted results.

3.2 Model Design

3.2.1 Overall Architecture

The proposed matching model of traffic sign image and text description adopts a dual-stream network architecture, which is mainly composed of two core components: image encoder and text encoder. Image Encoder: Responsible for extracting high-level visual features from traffic sign images. Text Encoder: Responsible for extracting deep semantic features from the text descriptions of traffic signs. The outputs of the two encoders (i.e., image embedding and text embedding) are first mapped to a shared multimodal embedding space () via their own independent linear projection layers. After projection, these embedding vectors are Layer Normalized and then L2 normalized to ensure that the length of each eigenvector is in units. This normalization process facilitates the subsequent use of Euclidean distance for an effective similarity measure. The training goal of the model is to make the feature representations of matching image-text pairs as close as possible in a shared embedding space, while the feature representations of unmatched image-text pairs are as far away as possible.

3.2.2 Image Feature Extraction

Three pre-trained image models were introduced to evaluate the performance of different backbone networks. After removing the classification head, the 2048-dimensional (the first two) or 1792-dimensional (latter) eigenvectors were extracted from them, and the unified dimensions were mapped into a unified dimension by linear layers. After that, layer, L2 normalization were carried out to obtain the final image embedding.

3.2.3 Text Feature Extraction

The text encoding part adopts three mainstream Transformer language models: BERT, RoBERTa, and DistilBERT. The hidden state of the [CLS] mark is used as the representation of the whole sentence (768 dimensions), and the text embedding of a unified dimension is generated through linear mapping, layer normalization and L2 normalization consistent with the image processing process.

3.2.4 Loss Function Design

The model was trained using a Contrastive Loss function based on Euclidean distance, combined with the In-batch Negative Sampling strategy, to effectively improve the ability to match images and text. Specifically, for each training batch of size B, the image encoder and the text encoder output L2 normalised embedding vectors (dimensions B×D), respectively, which were used to construct the Euclidean distance matrix between the image and text. For matching image-text pairs (positive sample pairs), the loss function closes their position in the embedded space by minimizing their squared distance. On the other hand, for unmatched pairs of images (negative sample pairs), if their distance is less than the preset marginal value margin (the default setting is 1.0), there will be an additional penalty loss that pushes them away from the minimum margin. The loss function is optimised by simultaneously using all positive and negative sample pairs in each batch so that the model can efficiently learn a multimodal embedding space with intense discrimination and accurate matching.

Table6

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Batch Negative Sampling Significantly Improves Retrieval Performance

As shown in the table, in the experiment with the batch size set to 32 and 80 rounds of training, the In-Batch Negative Sampling strategy is significantly better than the Random Negative Sampling. The retrieval performance of almost all model combinations has been dramatically improved, and R@1 has generally jumped from dozens or even single digits to more than 90%.

4.2 Comparison of Different Encoder Combinations

Based on the average performance of batch sizes 32 and 64, we comprehensively compare nine text-to-

Table 2: *Retrieval Performance of Different Model Combinations under Sampling Strategies.*

(Batch Size = 32, Epoch = 80)		Random Negative Sampling					In Batch Negative Sampling				
Text	Image	R@1	R@5	R@10	MRR	MAP	R@1	R@5	R@10	MRR	MAP
Distilbert	Resnet50	21.69	88.49	99.60	46.82	46.81	92.82	99.87	99.87	96.35	96.35
RoBerta	Resnet50	0.2	1.12	2.31	1.52	1.52	46.49	99.87	99.87	71.78	71.78
Bert	Resnet50	26.59	77.61	90.94	48.72	48.72	92.72	99.87	99.90	96.29	96.29
Distilbert	Efficientnet_B4	21.86	90.24	99.77	47.52	47.51	95.11	99.87	99.87	97.48	97.48
RoBerta	Efficientnet_B4	1.42	8.80	17.29	7.34	7.34	43.06	98.74	99.77	68.62	68.62
Bert	Efficientnet_B4	37.93	96.06	99.01	62.82	62.82	94.01	99.87	99.87	96.93	96.93
Distilbert	Resnext50_32x4d	35.95	94.61	99.93	61.13	61.13	93.32	99.87	99.87	96.57	96.57
RoBerta	Resnext50_32x4d	0.13	0.50	0.86	0.49	0.34	24.14	97.06	99.93	50.73	50.73
Bert	Resnext50_32x4d	19.08	57.71	75.03	37.00	37.00	47.62	99.87	99.87	72.61	72.61

Table 3: *Retrieval performance under In-Batch Negative Sampling strategy, Batch Size=32.*

(Batch Size = 32, Epoch = 80)		Text to image					Image to text				
Text	Image	R@1	R@5	R@10	MRR	MAP	R@1	R@5	R@10	MRR	MAP
Distilbert	Resnet50	92.86	99.87	99.87	96.37	96.37	92.53	99.87	99.87	96.19	96.19
RoBerta	Resnet50	46.36	99.87	99.87	71.69	71.69	46.03	99.87	99.87	71.40	71.40
Bert	Resnet50	93.52	99.87	99.93	96.69	96.69	92.06	99.87	99.87	95.94	95.94
Distilbert	Efficientnet_B4	95.50	99.87	99.87	97.69	97.69	94.71	99.87	99.87	97.28	97.28
RoBerta	Efficientnet_B4	44.31	98.61	99.80	69.22	69.22	41.73	99.07	99.74	68.05	68.05
Bert	Efficientnet_B4	94.18	99.87	99.87	97.03	97.03	93.78	99.87	99.87	97.03	97.03
Distilbert	Resnext50_32x4d	93.72	99.87	99.87	96.79	96.79	93.06	99.87	99.87	96.43	96.43
RoBerta	Resnext50_32x4d	23.74	97.02	99.93	50.56	50.56	23.74	96.63	99.93	50.26	50.26
Bert	Resnext50_32x4d	47.29	99.87	99.87	72.55	72.55	46.96	99.87	99.87	72.17	72.17

Table 4: *Retrieval performance under In-Batch Negative Sampling Strategy, Batch Size=64.*

(Batch Size=64, Epoch=80)		Text to image					Image to text				
Text	Image	R@1	R@5	R@10	MRR	MAP	R@1	R@5	R@10	MRR	MAP
Distilbert	Resnet50	94.91	99.87	99.87	97.38	97.38	94.91	99.87	99.87	97.37	97.37
RoBerta	Resnet50	94.31	99.87	99.93	97.10	97.10	94.11	99.87	99.87	97.00	97.00
Bert	Resnet50	95.50	99.87	99.87	97.69	97.69	94.97	99.87	99.87	97.43	97.43
Distilbert	Efficientnet_B4	97.02	99.87	99.87	98.45	98.45	96.76	99.87	99.87	98.31	98.31
RoBerta	Efficientnet_B4	96.83	99.87	99.87	98.35	98.35	96.36	99.87	99.87	98.12	98.12
Bert	Efficientnet_B4	96.56	99.87	99.87	98.22	98.22	96.30	99.87	99.87	98.09	98.09
Distilbert	Resnext50_32x4d	97.62	99.87	99.87	98.75	98.75	97.29	99.87	99.87	98.57	98.57
RoBerta	Resnext50_32x4d	24.27	98.54	99.93	51.19	51.19	25.53	98.41	99.93	52.25	52.25
Bert	Resnext50_32x4d	96.76	99.87	99.87	98.32	98.32	95.97	99.87	99.87	97.92	97.92

Table 5: *Retrieval performance under In-Batch Negative Sampling strategy, Batch Size 32&64-average.*

(32&64-averageepoch=80)		32					64				
Text	Image	R@1	R@5	R@10	MRR	MAP	R@1	R@5	R@10	MRR	MAP
Distilbert	Resnet50	92.82	99.87	99.87	96.35	96.35	95.07	99.87	99.87	97.46	97.46
RoBerta	Resnet50	46.49	99.87	99.87	71.78	71.78	94.08	99.87	99.90	96.98	96.98
Bert	Resnet50	92.72	99.87	99.90	96.29	96.29	95.24	99.87	99.90	97.56	97.56
Distilbert	Efficientnet_B4	95.11	99.87	99.87	97.48	97.48	96.99	99.87	99.87	98.43	98.43
RoBerta	Efficientnet_B4	43.06	98.74	99.77	68.62	68.62	96.33	99.87	99.87	98.10	98.10
Bert	Efficientnet_B4	94.01	99.87	99.87	96.93	96.93	96.49	99.87	99.87	98.19	98.19
Distilbert	Resnext50_32x4d	93.32	99.87	99.87	96.57	96.57	97.49	99.87	99.90	98.68	98.68
RoBerta	Resnext50_32x4d	24.14	97.06	99.93	50.73	50.73	25.66	98.35	99.93	52.20	52.20
Bert	Resnext50_32x4d	47.62	99.87	99.87	72.61	72.61	96.20	99.87	99.87	98.03	98.03

image encoder combinations, and batch size 64 completely outperforms batch size 32.

4.3 Detailed Analysis of Text and Image Retrieval Results

As shown in Table 6, input a specific image query, and the top 5 text retrieval results will be displayed. The number on the right side of each text indicates the matching score, which means the similarity between the image and the text description. Most of the results show that our model can successfully retrieve a specific scene's text description.

Of course, our model also shows certain limitations. As shown in the figure, we have listed some typical errors. In Group A, although results A3, A4, and A5 did not query the correct road sign information, the road conditions, where the road signs are, whether there are obstacles, and the shape and colour of the road signs were correct. However, it confused "speed bumps", "hump bridges ahead", and "rolling stones on the left" with "uneven roads ahead", mainly because the image details are too detailed to distinguish. This proof further shows that although the model will confuse similar information related to the last three, it can also work well in the first two and clearly describe what happened. In group B, the content of the road signs in the image is not as easy to confuse as in group A. Some examples consider a straight and right turn at a multi-level crossing as straight and right turn, while others place the road signs on the wrong side. In group C, most examples of errors are that the information on the road signs is correct, but the identification of obstacles is wrong.

As shown in Table 8, inputting a specific text query will display the top 5 image retrieval results. The first one is able to be accurately retrieved, and the subsequent ranking will have retrieval errors on which side the road sign is on and whether there are obstacles, but the content of the road signs is mostly correct.

4.4 Conclusions and Practical Recommendations

- Text Encoder Selection: It is paramount to choose a powerful text encoder. Advanced models such as DistilBERT or BERT are prerequisites for high retrieval accuracy, overshadowing the importance of the image encoder in this context.
- Best Combinations: The pairs DistilBERT+EfficientNet_B4 and BERT+EfficientNet_B4 are recommended as the top-performing combinations, suitable for scenarios where high precision is critical.
- Resource-Constrained Scenarios: The DistilBERT+ResNet50 combination balances between performance and efficiency, making it ideal for applications with limited resources.
- Batch Size Considerations: For most DistilBERT combinations, a smaller batch size (32) is sufficient. In contrast, combinations like BERT+ResNeXt50_32x4d and all RoBERTa pairings benefit significantly from larger batch sizes to achieve better performance.
- Combinations to Avoid: Any combination involving RoBERTa performs relatively weakly, especially RoBERTa+ResNeXt50_32x4d, and should be avoided in practical applications.

These findings offer valuable guidance for model selection in traffic sign image and text description matching tasks, helping to strike the optimal balance between performance and efficiency:

Table 6: *Typical results of text retrieval from an image and its similarity to the query result identified.*

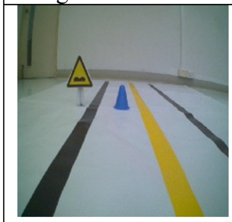
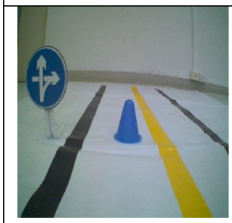
Image	Captions	Accuracy
 "image_path":" xy_049_049_ f5333332-f67a- 11ef-a814- 00e04c4bcba9.jpg"	"true caption": "A triangular sign, the 'Rough Surface Ahead' sign, yellow-colored with a black outline, is found on the left side of a straight road, as the sign depicts dips and humps, indicating potential jolts or reduced stability on the road. Moreover, an obstacle is set in the center.",	
	"top1 captions": "A triangular sign, the 'Rough Surface Ahead' sign, yellow-colored with a black outline, is found on the left side of a straight road, as it the sign depicts dips and humps, indicating potential jolts or reduced stability on the road. Moreover, an obstacle is set in the center."	94.72
	"top2 captions": "On a straight road, installed on the left side, you can see a triangular sign, the sign named 'Dip', yellow-colored with a black outline, where it shows a road dip shape, warning of a sudden low section. Meanwhile, a roadblock is present."	70.93
	"top3 captions": "A triangular sign of the 'Speed Bump' warning sign, yellow-colored with a black outline, is found on the left side of a linear road, as the sign warns drivers of a bump in the road requiring reduced speed. Moreover, there is a barrier in the middle.",	58.15
	"top4 captions": "On a linear road, you will notice a triangle-shaped warning sign, the 'Hump Bridge Ahead' sign with a yellow background and black border installed on the left side, which it shows a hump shape, indicating a steep bridge that may obstruct sight while a roadblock is present."	56.80
	"top5 captions": "Observing a direct roadway from a position on the left side, one finds a triangle-shaped sign, the 'Falling Rocks (Left Side)' sign with a yellow background and black border that depicts debris tumbling down from the left side. Plus, a roadblock is present."	56.55
 "image_path": " xy_050_050_ a0a7c156-f680 -11ef-89b7- 00e04c4bcba9.jpg"	"true_caption": "A set installed on the left side of a straight road is a round sign, the 'Straight and right turn at multi-level crossing' sign with a blue background; it shows an upward arrow plus a right-turn arrow, with a barrier in the middle in place."	
	"top1_captions": "A set installed of the left side on a straight road is a round sign, the 'Straight and right turn at multi-level crossing' sign with a blue background; it shows an upward arrow plus a right-turn arrow, with a barrier in the middle in place."	96.89
	"top2_captions": "Along a direct roadway, on the left side displays a round mandatory sign, the sign indicating forward and right turns are allowed with a blue background and white arrows that it shows two arrows pointing straight and right, indicating vehicles may proceed forward or make a right turn; furthermore, a roadblock is present."	61.42
	"top3_captions": "Here is a T-style intersection, and on the left stands a circular sign, the sign named 'Elevated intersection: straight and right turn' with a blue background. Notably, it shows an upward arrow plus a right-turn arrow, with no obstacle in the center."	59.77

Table 6: *Typical results of text retrieval from an image and its similarity to the query result identified Cont.*

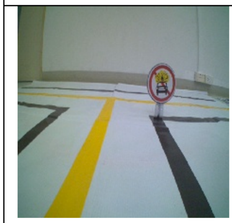
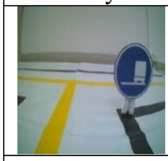
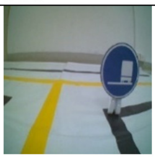
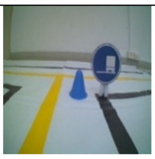
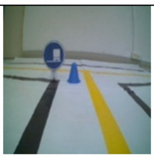
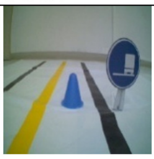
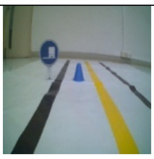
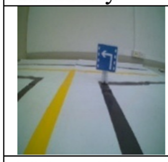
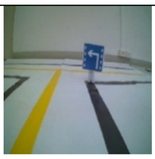
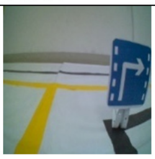
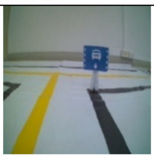
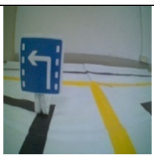
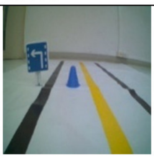
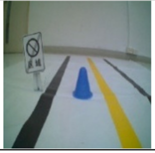
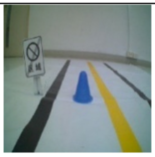
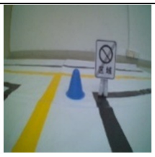
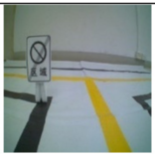
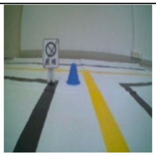
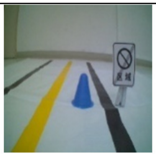
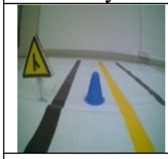
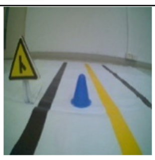
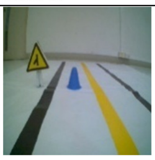
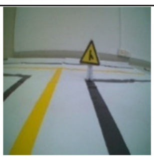
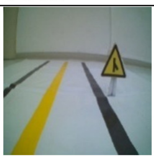
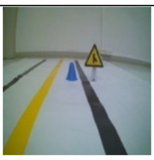
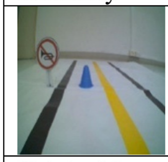
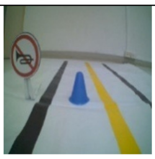
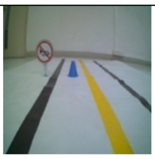
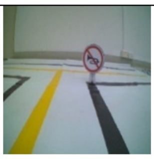
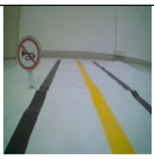
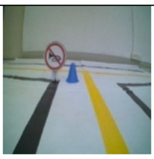
Image	Captions	Accuracy
	"top4_captions": "The 'Straight and right turn at multi-level crossing' sign, a circular sign and blue-colored, is positioned on the right side of a T-junction, where it shows an upward arrow plus a right-turn arrow. In addition, there is a barrier in the middle."	59.06
	"top5_captions": "On a direct roadway, you will notice a round sign, the sign named 'Elevated intersection: straight and right turn' is blue-colored on the right side, which depicts a straight arrow and a right arrow while an obstacle is set in the center."	58.67
 "image_path": " xy_049_049_ 467d945e-f67a -11ef-969f- 00e04c4bcba9.jpg"	"true_caption": "The 'Hazardous materials transport prohibited' sign, a circular sign featuring a red border on white, is positioned on the right side of a T-style intersection, signifying that it forbids vehicles carrying hazardous materials from entering. In addition, no roadblock is placed.",	
	"top1_captions": "The 'Hazardous materials transport prohibited' sign, a circular sign featuring a red border on white, is positioned on the right side of a T-style intersection, signifying that it forbids vehicles carrying hazardous materials from entering. In addition, no roadblock is placed."	96.96
	"top2_captions": "The sign named 'No dangerous goods vehicles' is a circular sign featuring a red border on white, located on the right side of a T-junction, meaning the center shows an exploding icon above a car silhouette crossed out by a slash. Also, there is a barrier in the middle."	60.94
	"top3_captions": "On a linear road, you will notice a round sign, the sign named 'No dangerous goods vehicles' featuring a red border on white placed on the right side, clarifying that the center shows an exploding icon above a car silhouette crossed out by a slash while there is no obstacle in the center."	58.87
	"top4_captions": "A circular sign, the sign named 'No dangerous goods vehicles' featuring a red border on white, is found installed on the right side of a straight road, declaring the center shows an exploding icon above a car silhouette crossed out by a slash. Moreover, there is a barrier in the middle."	58.26
	"top5_captions": "Set on the left side of a straight road is a circular sign, the sign named 'No dangerous goods vehicles' with a red ring and white background; it conveys that it forbids vehicles carrying hazardous materials from entering, with no roadblock in place."	57.97

Table 7: *Traffic sign information retrieval results of actually retrieving text from images.*

Search task	Road type	Side	Obstacle	Sign name	Sign shape	Sign color	Sign content
A1	√	√	√	√	√	√	√
A2	√	√	√	√	√	√	√
A3	√	√	√	×	√	√	×
A4	√	√	√	×	√	√	×
A5	√	√	√	×	√	√	×
B1	√	√	√	√	√	√	√
B2	√	√	√	×	√	√	×
B3	√	√	×	√	√	√	√
B4	×	×	√	√	√	√	√
B5	√	×	√	√	√	√	√
C1	√	√	√	√	√	√	√
C2	√	√	×	√	√	√	√
C3	×	√	√	√	√	√	√
C4	×	√	×	√	√	√	√
C5	×	×	√	√	√	√	√

Table 8: *Typical results of image retrieval from a text and its similarity to the query result identified.*

	Top1	Top2	Top3	Top4	Top5
Accuracy	100	85.71	71.42	71.42	57.14
 Original Image					
Input caption	"On a T-style intersection, you will notice a round mandatory sign, the 'Keep Right' sign blue with white symbols showing a vehicle on the right installed on the right side, which the sign mandates vehicles to keep to the right lane unless overtaking is required while there is no obstacle in the center."				
Accuracy	100	71.42	71.42	85.71	85.71
 Original Image					
Input caption	"Observing a T-junction from on the right side, one finds a rectangular sign, the 'Left-turn Lane' sign featuring a white left arrow on a blue field that denotes the center shows a bold left-turn arrow. Plus, no roadblock is placed."				
Accuracy	100	85.71	71.42	71.42	85.71
 Original Image					
Input caption	"On a straight road, you will notice a vertical rectangle sign, the sign named 'Area no long parking lifted' with a white base showing the end of the no-long-parking zone on the left side, declaring that the diagonal slash denotes the end of the no-long-parking restriction while there is a barrier in the middle."				
Accuracy	100	85.71	57.14	71.42	85.71
 Original Image					
Input caption	"Here is a direct roadway, and placed on the left side stands a triangular sign named 'Intersection Ahead' with a yellow background and black border. Notably, it shows a left-side branch intersection, and a roadblock is present."				
Accuracy	100	100	57.14	85.71	85.71
 Original Image					
Input caption	"A round sign named 'No honking' featuring a red border on white is found on the left side of a direct roadway, declaring it forbids vehicles from using their horn. Moreover, an obstacle is set in the center."				

5. CONCLUSION AND FUTURE WORK

This paper proposes a method for matching traffic sign images and text descriptions based on deep learning. This method parses the directory structure of the dataset to automatically generate fine-grained category labels, reducing the cost of manual annotation. At the same time, an InfoNCE loss function based on negative sampling within the batch is designed to fully use the non-matching sample pairs in each training batch, thereby improving the model's discriminative ability. After experiments, the combination of Resnext50_32x4d and DistilBERT achieved a high accuracy rate. This method has a good application in the future in autonomous driving scenarios and can help intelligent driving systems more accurately identify and understand traffic signs and improve driving safety. Although there is a significant improvement, problems exist, such as easy confusion between similar signs and a less-than-ideal data environment. Future research will focus on expanding the dataset in complex environments and exploring multimodal fusion and more advanced models to further enhance the practicality and robustness of the model.

The source code and dataset for this study can be found at `https://github.com/William990313/Research-on-Deep-Learning-Based-Methods-for-Matching-Traffic-Sign-Images-with-Textual-Descriptions.git`.

ACKNOWLEDGEMENT

The first author was awarded the CPALL scholarship while conducting this research in PIM.

AUTHOR CONTRIBUTIONS

Conceptualization, N.W. and J. Q.; methodology. N.W. and J. Q.; software, N.W. and J. Q.; validation. N.W. and J. Q.; formal analysis, N.W. and J. Q.; investigation, N.W. and J. Q.; data curation, N.W.and J. Q.; writing—original draft preparation, N.W.and J. Q.; writing—review and editing, N.W. and J.Q.; visualization, N.W.and J. Q.; supervision, J. Q.; J.Q.is the corresponding author.; All authors have read and agreed to the published version of the manuscript.

References

[1] Y. Li and J. Qu, "Intelligent Road Tracking and Real-time Acceleration-Deceleration for Autonomous Driving Using Modified Convolutional Neural Networks," *Current Applied Science and Technology*, vol. 22, no. 1, pp. 1–10, 2022.

[2] T. Chen, S. Kornblith, M. Norouzi, & G. Hinton, "UniT: Multimodal multitask learning with a unified transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1439–1449, 2021.

[3] L. Yang, P. Luo, C. C. Loy, & X. Tang, "Fine-grained traffic sign recognition with hierarchical attention and localization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 5967–5979, 2022.

[4] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, & A. Torralba, "Learning to generate fine-grained image labels with minimal supervision," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, pp. 3576–3584, 2022.

[5] K. He, H. Fan, Y. Wu, S. Xie, & R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020.

[6] A. van den Oord, Y. Li, & O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[7] Y. Xie & J. Qu, "A study on bilingual deep learning PIS neural network model based on graph-text modal fusion," *ECTI Transactions on Computer and Information Technology*, vol. 19, no. 1, pp. 13–24, 2025.

[8] F. Zheng & J. Qu, "TIDCB: Text image dangerous-scene convolutional baseline," *ECTI Transactions on Computer and Information Technology*, vol. 18, no. 3, pp. 45–54, 2024.

[9] Y. Xie & J. Qu, "A study on Chinese language cross-modal pedestrian image information retrieval," *Songklanakarin Journal of Science and Technology*, vol. 46, no. 5, pp. 466–475, 2024.

[10] Z. Liu *et al.*, "Traffic sign captioning: Generating contextual descriptions for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1567–1581, 2023.

[11] Y. Wang *et al.*, "AutoLabel: Weakly supervised learning for efficient traffic sign annotation," *Pattern Recognition*, vol. 135, pp. 109–123, 2023.

[12] X. Chen *et al.*, "Fine-grained traffic sign recognition with attribute-guided attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10245–10254, 2022.

[13] R. Zhang *et al.*, "Contrastive learning for cross-modal traffic sign retrieval," *Engineering Applications of Artificial Intelligence*, vol. 129, pp. 107–120, 2024.

[14] N. Wang & J. Qu, "Explainable image captioning for autonomous driving: A traffic sign recognition task," in *Proceedings of the IEEE International Conference on Business and Industrial Research (ICBIR)*, to be printed in IEEE Explore, 2025.

[15] Z. Zhao *et al.*, "Masking-Based Cross-Modal Remote Sensing Image–Text Retrieval via Dynamic Contrastive Learning," in *IEEE Transactions on*

Geoscience and Remote Sensing, vol. 62, pp. 1-15, 2024.

Ning Wang is currently studying for the Master of Engineering Technology, Faculty of Engineering and Technology, Panyapiwat Institute of Management, Thailand. He received B.A. Architect from Nanjing Tech University Pujiang Institute, China, in 2023. His research interests are Research direction is artificial intelligence, image processing, and natural language processing (NLP). He is awarded with a full scholarship from CPALL while conducting his research in PIM.

Jian Qu is an Assistant professor at the Faculty of Engineering and Technology, Panyapiwat Institute of Management. He received Ph.D. with Out Standing Performance award from Japan Advanced Institute of Science and Technology, Japan, in 2013. He received B.B.A with Summa Cum Laude honors from Institute of International Studies of Ramkhamhaeng University, Thailand, in 2006, and M.S.I.T from Sirindhorn International Institute of Technology, Thammast University, Thailand, in 2010. He has been a house committee for Thai SuperAI since 2020. His research interests are natural language processing, intelligent algorithms, machine learning, machine translation, information retrieval and image processing.