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

TIDCB: Text Image Dangerous-Scene Convolutional Baseline

Fangfang Zheng[1] and Jian Qu[2]

ABSTRACT

The automatic management of public area safety is one of the most challenging issues in our society. For instance, the timely evacuation of the public during incidents such as fires or large-scale shootings is paramount. However, detecting pedestrian behavior indicative of danger promptly from extensive video surveillance data may not always be feasible. This may result in untimely warnings being provided, resulting in significant loss of life. Although existing research has proposed “text-based person search,” it has primarily focused on pedestrian search by matching images of pedestrian body parts to text, lacking the search for pedestrians in dangerous scenarios. To address this gap, this paper proposes an innovative warning framework that further searches for individuals in hazardous situations based on textual descriptions, aiming to prevent or mitigate crisis events. We have constructed a new public safety dataset named CHUK-PEDES-DANGER, one of the first pedestrian datasets that includes dangerous scenes. Additionally, we introduce a novel framework for public automatic evacuation. This framework leverages a multimodal deep learning architecture that combines the image model ResNet-50 with the text model RoBERTa to produce our Text-Image Dangerous-Scene Convolutional Baseline (TIDCB) model, which addresses the classification problem from text to image and image to text by matching images of pedestrian body parts and environments to text. We propose a novel loss function, cross-modal projection matching-triplet (CMPM-Triplet). After conducting extensive experiments, we have validated that our method significantly improves accuracy. Our model outperforms TIPCB with a matching rate of 76.93%, an improvement of 4.78% compared to TIPCB, and demonstrates significant advantages in handling complex scenarios.

Article information:

Keywords: Cross-modality, Image-to-Text, Person Search, Dangerous Scene Search, Early Warning Framework

Article history:
Received: January 4, 2024
Revised: March 11, 2024
Accepted: May 9, 2024
Published: May 25, 2024
(Online)

DOI: 10.37936/ecti-cit.2024183.255314

1. INTRODUCTION

In the current world, there is a growing need for advanced surveillance systems as civilization develops. This is particularly true when processing extensive video data to target vehicles or pursue criminals. In these situations, traditional approaches may need weeks or even months of police effort, significantly depleting societal resources. Furthermore, as the public becomes more aware of legal rights, worries about privacy and portrait rights may surface.

Re-identification (ReID)[1] technology has surfaced as an alternative to facial recognition in light of these difficulties. ReID allows a person to be identified and tracked within video data by using descriptions of images or attributes. However, in some situations, like following criminal suspects in criminal investigations, acquiring clear images of people can be very challenging. In light of this, a new pedestrian search sub-domain based on natural language descriptions has emerged: the Text-Image Part-based Convolutional Baseline (TIPCB)[2]. TIPCB significantly improves retrieval efficiency by integrating multi-modal information to enable effective cross-modal retrieval. In addition to advancing ReID technology, this approach has dramatically increased the efficiency of intelligent monitoring in real-world situations.

[1,2] The authors are with the Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi, Thailand. E-mail: 6572100031@stu.pim.ac.th and jianqu@pim.ac.th
[2] Corresponding author: jianqu@pim.ac.th

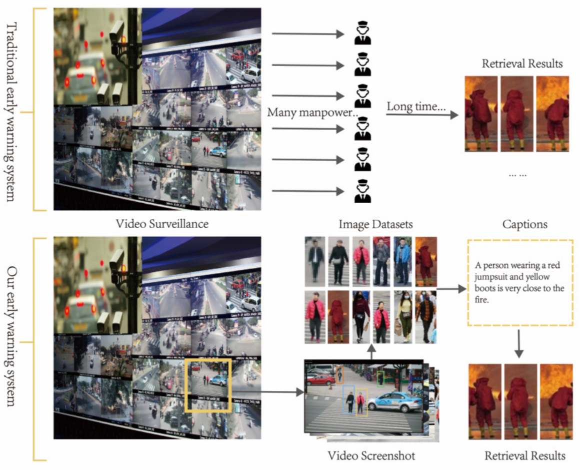


Fig.1: *Comparison of traditional and our early warning system.*

Despite the exceptional work done by Li, Shuang [3] and Han, Xiao[4] *et al.* in "Person Search Based on Natural Language," we think there is still a lot more to explore. The previous research was limited to searching for individual characteristics of persons and did not involve searching for individuals in dangerous scenarios. For instance, we might not be able to quickly identify pedestrian dangers in crowded public areas like streets, shopping malls, and communities from extensive video surveillance data. For example, a large fire occurred at an outlet in Daejeon, South Korea[5], on September 26, 2022, which resulted in seven fatalities and one serious injury. In addition, a shooting at the Siam Paragon in Bangkok on October 3, 2023[6], resulted in three deaths and five wounded. Perhaps such accidents could be avoided, or the number of victims could be decreased if we could use surveillance to identify the location of a suspect or provide early warnings when a fire breaks out. As shown in Fig 1, the traditional early warning system relies on intelligent surveillance technology, requiring security personnel to constantly monitor for unusual activities in surveillance footage, such as fires or shootings. Given the vast surveillance data, relying solely on human observation poses risks of oversight and manpower shortage.

This made us wonder if we can use text to search for people in dangerous scenarios in the future. Thus, we proposed a novel early warning framework, expanding the current cross-modal text-to-image retrieval framework. We construct various images of individuals captured from surveillance footage, each accompanied by corresponding textual annotations. Unlike previous cross-modal text-to-image approaches, our annotations describe individual characteristics and highlight features indicative of potential danger in the scene. Thus, we can link the environmental image feature to text annotations. Through these enhancements, the early warning system will be able to more accurately identify and alert to hazardous behaviors, enabling timely intervention in emergencies and reducing potential harm.

The first challenge we encountered was a need for more datasets. Gathering comprehensive statistics on pedestrians is complex, and exploring the retrieval of dangerous scenarios is a particularly rare dataset. Second, integrating this innovative search approach into existing research has some challenges. To identify a pedestrian, existing works extract features only from several pedestrian body parts (head, shoulders, body, legs, and clothes) to connect text and picture attributes; they ignore environmental information to boost retrieval accuracy. Now, we need to extend such an idea to include both pedestrian body parts and environmental features to make the connection between text and picture attributes. The added environmental feature in the image could sometimes overlay with body part features. These modifications could harm the retrieval accuracy.

To address the lack of dataset issues, we chose to modify images from the CUHK-PEDES dataset to produce the CHUK-PEDES-DANGER dataset. With each image labeled, this simulates images of people in dangerous situations, including "fires," "floods," and "crossing through bushes," addressing the existing shortage of large-scale datasets to assist this emerging field of study. Furthermore, we changed the language model in TIPCB to the RoBERTa[7] language model and proposed a novel loss function called CMPM-Triplet. Experiments show that the newly introduced function effectively improves model performance.

The main contributions of this paper are: (1) A novel early warning framework based on individual searches in dangerous situations and descriptions in natural language is proposed. (2) To address the scarcity of extensive datasets to enable the emerging field of "natural language search for people in dangerous scenes," the CHUK-PEDES-DANGER public safety dataset was developed. (3) We created our model TIDCB by modifying the TIPCB model, and we also introduced a novel loss function called CMPM-Triplet. (4) Our approach outperforms the TIPCB model based on numerous experiments performed on the CHUK-PEDES and CHUK-PEDES-DANGER datasets.

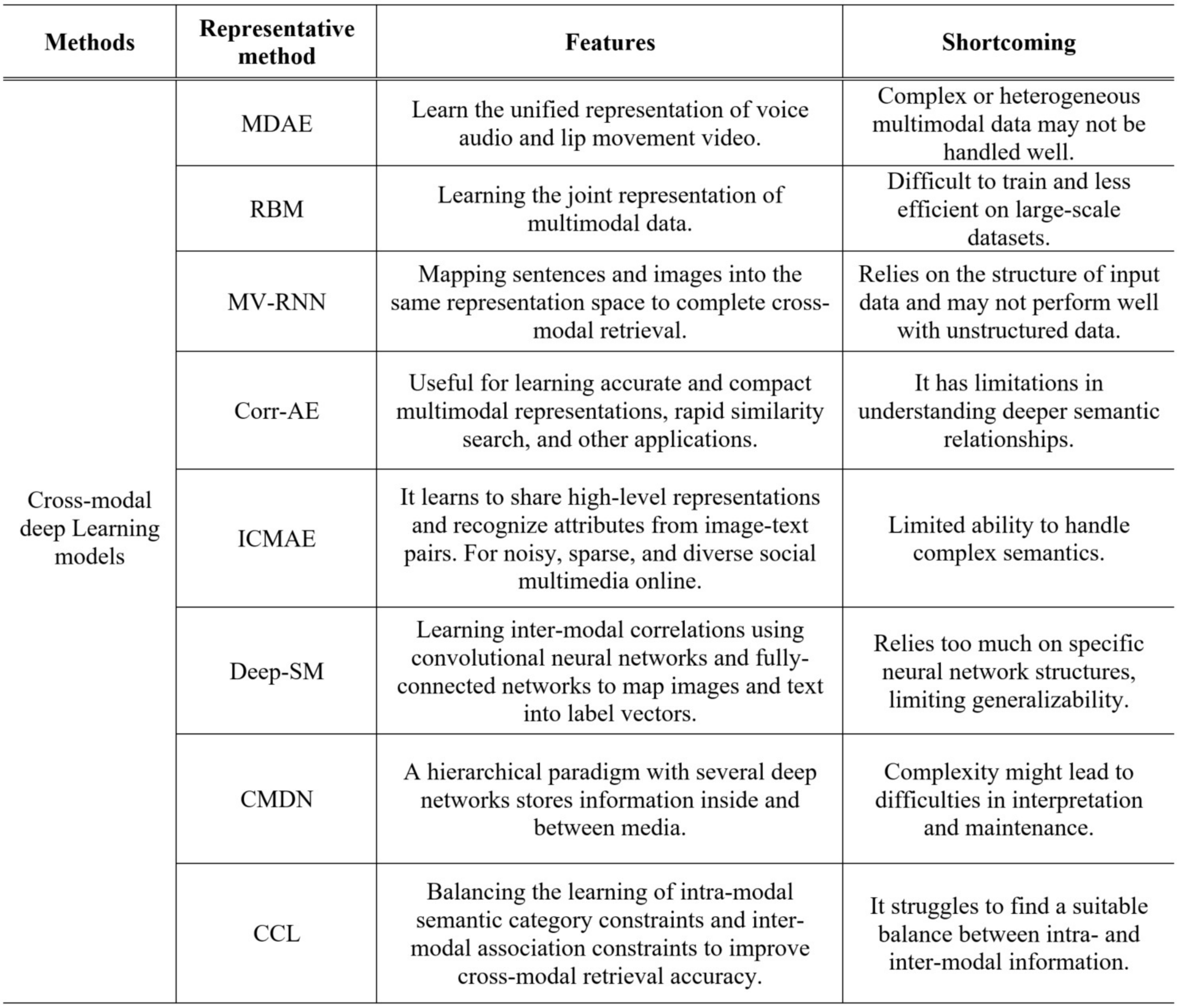
2. RELATED WORK

2.1 Research on dangerous scene detection

Dangerous scene detection has always been essential in public safety and social management. Quickly and accurately recognizing potential threats in pictures or surveillance videos is crucial for preventing accidents and criminal activities.

In response to this issue, H. Yang *et al.* [8] proposed a real-time framework for detecting dangerous

Table 1: *Comparison of Cross-Modal Retrieval Methods in Person Search with Natural Language Description.*



Methods	Representative method	Features	Shortcoming
Cross-modal deep Learning models	MDAE	Learn the unified representation of voice audio and lip movement video.	Complex or heterogeneous multimodal data may not be handled well.
	RBM	Learning the joint representation of multimodal data.	Difficult to train and less efficient on large-scale datasets.
	MV-RNN	Mapping sentences and images into the same representation space to complete cross-modal retrieval.	Relies on the structure of input data and may not perform well with unstructured data.
	Corr-AE	Useful for learning accurate and compact multimodal representations, rapid similarity search, and other applications.	It has limitations in understanding deeper semantic relationships.
	ICMAE	It learns to share high-level representations and recognize attributes from image-text pairs. For noisy, sparse, and diverse social multimedia online.	Limited ability to handle complex semantics.
	Deep-SM	Learning inter-modal correlations using convolutional neural networks and fully-connected networks to map images and text into label vectors.	Relies too much on specific neural network structures, limiting generalizability.
	CMDN	A hierarchical paradigm with several deep networks stores information inside and between media.	Complexity might lead to difficulties in interpretation and maintenance.
	CCL	Balancing the learning of intra-modal semantic category constraints and inter-modal association constraints to improve cross-modal retrieval accuracy.	It struggles to find a suitable balance between intra- and inter-modal information.

behaviors. This framework aims to identify the hazardous behaviors of workers in complex industrial settings. The authors conducted data analysis of actual production scenarios and decomposed the task requirements into functionalities such as object detection, scene recognition, behavior analysis, and safety logic reasoning. However, such a method is task-specific; thus, it only works for their experimental factory, and everything must be re-collected and redesigned if applied to a different area.

Ravanbakhshd *et al.*[9] proposed a novel paradigm for abnormal event detection by combining multimodal learning and Generative Adversarial Networks (GANs). They presented a GAN-based technique for anomaly detection in crowded settings. To teach GANs the internal representation of scene normalcy, normal video frames and the related optical flow images were used for training. GANs cannot produce anomalous occurrences because they are trained solely on normal data (without abnormal action). In the testing process, the appearance and motion representations reconstructed by the GANs are compared to real data, and abnormal actions are identified by computing local differences.

For the above methods, we noted three major issues with their approach: first, there is a shortage of clear and objective definitions of abnormal behavior; second, their design is problem specific, thus applying such design elsewhere will need a re-hull of the design; third, there are few datasets with ground truth for abnormal samples available, which is a major limitation for data-hungry deep learning methods. These three problems are related because creating large-scale annotated datasets is more difficult due to the subjectivity in defining abnormalities.

2.2 Cross-modal retrieval methods

With the popularity of deep learning in recent years, multimodal learning has also drawn a lot of interest, as demonstrated by Table 1. A self-encoder model (MDAE) was proposed by Ngiam *et al.* [10] to learn a unified representation of lip-activated video and speech audio. A depth-constrained Boltzmann machine (RBM) was created by Srivastava and Salakhutdinov[11] to learn a joint representation of multimodal data. Socher *et al.* [12] suggested a recurrent neural network model (MV-RNN) based on dependency trees to map phrases and images into the same space to achieve cross-modal retrieval. Self-encoders were used by Feng *et al.* [13] and Wang W. *et al.* [14] to build a cross-modal retrieval model (Corr-AE). To efficiently search for similarities in multimodal data and other relevant applications, Wang D. *et al.* [15] introduced a multimodal modal deep learning system that learns accurate and compact multimodal representations of multimodal data. Zhang *et al.* [16] proposed an attribute discovery method called Independent Component Multimodal Self-Encoder (ICMAE), which learns a shared high-level representation to recognize attributes from a set of image and text pairs. Aiming at the characteristics of web social multimedia content, which is noisy, sparse, and diverse, Zhang *et al.* [17] further proposed a method to learn a unified image-text representation from web social multimedia content. To learn inter-modal correlation, Wei *et al.* [18] suggested a deep semantic matching (deep-sm) method that maps text and images into labeled vectors using convolutional neural networks and fully connected networks. To retain information within and between media, Peng *et al.*[19]presented the CMDN model, a hierarchical framework with numerous deep networks. Additionally, they proposed the cross-modal correlation learning method (CCL)[20], which balances learning inter-modal correlation constraints with maintaining fine-grained information of various modalities while modeling coarse-grained information of various modalities. This approach preserves fine-grained information of various modalities while modeling the coarse-grained information of various modalities.

However, they all generally experience the following issues:(1) In deep learning-based techniques, current methodologies may only apply to some application scenarios, such as textual descriptions-based character searches, because they are tailored for particular kinds of multimodal data. (2) Working with novel and untested data kinds or structures may present challenges for these approaches.

In addition, similar challenges have emerged in other Natural language Processing domains. For instance, J. Qu and A. Shimazu [21] proposed automatic evaluation for cross-language information extraction and out-of-vocabulary (OOV) term translation. While this method has shown promising results in their respective fields, it still needs help with small dataset sizes and subjective definition ambiguity.

2.3 Text-based person search methods

In the Text-based Person Search (TBPS) field, Li *et al.* [3] conducted pioneering research on the challenging dataset CUHK-PEDES and constructed a baseline model based on a recurrent neural network with gated neural attention. Subsequently, various single-scale methods have been proposed, leveraging techniques such as instance loss [22], cross-modal projection loss [23], adversarial loss [24], and cross-modal knowledge adaptation[25] to better explore the intra- and inter-modal fine-grained differences.

One of the most outstanding research contributions is the TIPCB proposed by Yuhao Chen *et al.* [2]; it resulted in the development of CUHK-PEDES, a large-scale person dataset, and a baseline model based on Recurrent Neural Networks with Gated Neural Attention (GNA-RNN). In this framework, the most pertinent images for situations such as intelligent surveillance are retrieved using the CMPM algorithm, which ranks all images in the dataset according to the relation between text descriptions and images.

We create a new public safety dataset called CHUK-PEDES-DANGER and propose an inventive early warning framework, building upon the original cross-modal retrieval paradigm by Yuhao Chen *et al.*, addressing the issue of data scarcity. To prevent or resolve crisis events, this dataset makes it possible to retrieve people from dangerous situations inside big image datasets. This has significant implications for public safety awareness.

3. METHODOLOGY

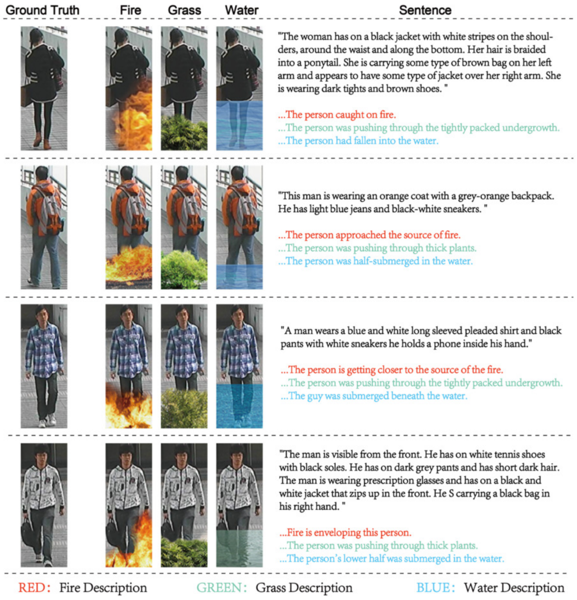
In the following chapters, we will discuss in Section 3.1 how we created the CHUK-PEDES-DANGER dataset containing dangerous scenarios. Section 3.2 describes our proposed CMPM-Triplet function and its application to a multimodal model. Finally, Section 3.3 will explain how we provide early real-life warning mechanisms through our new image-to-text and text-to-image search methods.

3.1 Dataset construction

We made a dataset called CUHK-PEDES-DANGER just for public safety. It is an improvement on the CUHK-PEDES dataset. The CUHK-PEDES dataset is constructed at the Chinese University of Hong Kong. It is primarily used for pedestrian searches based on natural language descriptions. It has over 40,000 images of 1,003 pedestrians, and each one has at least two written descriptions, for a total of over 80,000 descriptions; these descriptions encompass information regarding pedestrians' appearance features, clothing styles, behavioral actions, and other details.

To make the CUHK-PEDES-DANGER dataset, we took 2385 different pedestrian instances from

CUHK-PEDES and picked four images of each instance in a different pose. This gave us a total of 9540 original images. This dataset is split into two parts. The first part has the 9540 original images from CUHK-PEDES. The second part has 9540 images that have been modified, as seen in Fig. 2. Given the relative scarcity of large datasets for dangerous scenarios; we generated images of hazardous scenes using photo edit technology. We specifically added pre-selected elements like "water," "fire," and "grass" to the original pictures to make them look like dangerous situations similar to "floods," "fires," and illegal activity like "crossing through bushes." During this process, we used a variety of "water, fire, grass" elements to enhance the diversity of the dataset.



"The woman has on a black jacket with white stripes on the shoulders, around the waist and along the bottom. Her hair is braided into a ponytail. She is carrying some type of brown bag on her left arm and appears to have some type of jacket over her right arm. She is wearing dark tights and brown shoes. "

...The person caught on fire.
...The person was pushing through the tightly packed undergrowth.
...The person had fallen into the water.

"This man is wearing an orange coat with a grey-orange backpack. He has light blue jeans and black-white sneakers. "

...The person approached the source of fire.
...The person was pushing through thick plants.
...The person was half-submerged in the water.

"A man wears a blue and white long sleeved pleaded shirt and black pants with white sneakers he holds a phone inside his hand."

...The person is getting closer to the source of the fire.
...The person was pushing through the tightly packed undergrowth.
...The guy was submerged beneath the water.

"The man is visible from the front. He has on white tennis shoes with black soles. He has on dark grey pants and has short dark hair. The man is wearing prescription glasses and has on a black and white jacket that zips up in the front. He S carrying a black bag in his right hand. "

...Fire is enveloping this person.
...The person was pushing through thick plants.
...The person's lower half was submerged in the water.

Fig.2: *Samples in the CUHK-PEDES-DANGER dataset.*

Furthermore, we enhanced each image by including further textual explanations of the related hazardous situations, expanding upon the two existing descriptions written in natural language. These descriptions are quite detailed and encompass the individuals' physical appearance, behavior, stance, and interactions with other items. Typically, descriptions consist of 30 to 50 words.

To summarize, our recently created CUHK-PEDES-DANGER dataset comprises 2385 unique pedestrians, 19080 pictures, and 38160 textual descriptions.

3.2 Multimodal model and our loss function

In our proposed framework, shown in Fig. 3, three main parts comprise the image-text architecture. The first is the visual module, which uses the ResNet-50[26] and RoBERTa models to extract multimodal information. The second is the joint learning module, which coordinates and learns cross-modal representations. The third is a real-time early warning framework demonstration. At first, we preprocess data from different modalities. For example, textual data and image data are fed separately into their models to get text features and image features. When you give our model some text (t), it searches the image database for the image (i) that most closely matches the text (t). If t and i both have the same label, we have a good match. Using the TIDCB model, we get a shared representation of the data across modalities and make a label mask M_{ti} for each pair of samples. $M_{ti} = 1$ means that samples t and i have the same label, and $M_{ti} = 0$ implies that they have different labels. Then, we find the cosine similarity between the text and the picture. Finally, we sorted the candidates by score from highest to lowest to find the image with the highest similarity score.

3.2.1 Image feature extraction

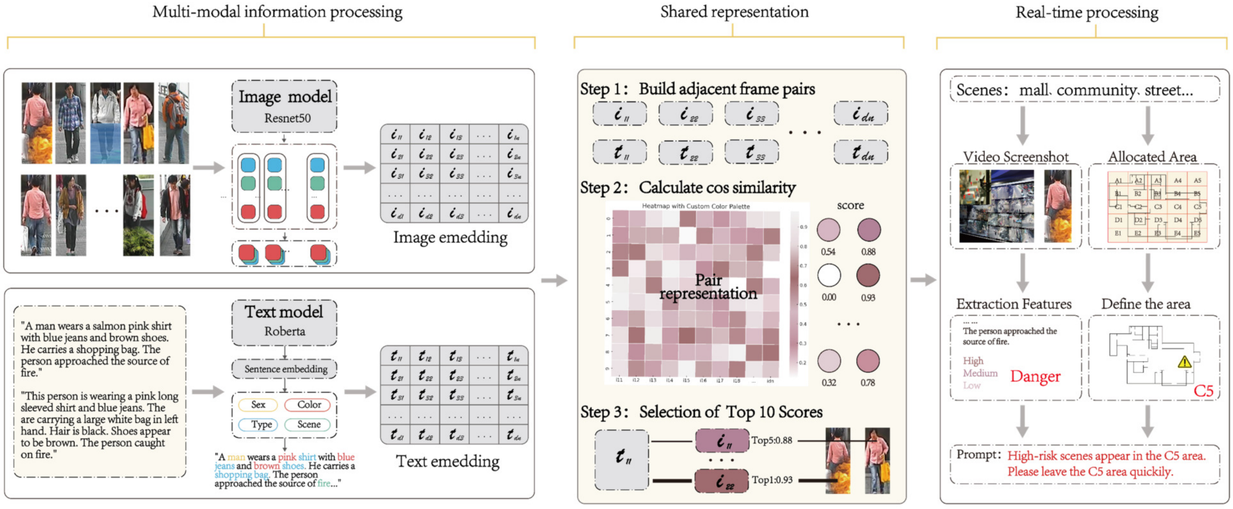
As shown in Fig. 4, the ResNet-50 network was the main design we used for the image feature extraction part of our study. ResNet-50 has four residual modules, and each one is made with residual links to help with network degradation. This design ensures maintained or enhanced performance even as network depth increases. These leftover blocks are very important for getting meaningful information from low to high levels, which makes features more expressive. In the last step, a matrix is made after the fully linked layers have been processed.

3.2.2 Text feature extraction

As shown in Fig. 5, we employed the RoBERTa model, a language representation model based on the Transformer[27]; it is a modification from BERT[28] for extracting text features. Specifically, the textual description t is first decomposed into a sequence of words, which are then transformed into a sequence of tokens through a pre-trained tokenizer. To handle texts of varying lengths uniformly, we adopted either a truncation or zero-padding strategy, resulting in a token sequence of length L. Subsequently, each token sequence is inputted into the RoBERTa model, which outputs a word embedding matrix of dimension D, where D represents the dimensionality of each word embedding. To align with the input requirements of the convolutional network layers, adjustments are made to the output measurements of the text embedding. Based on this, we constructed a multi-branch text convolutional neural network inspired by the principles of deep CNNs.

The main components of the RoBERTa model comprise multiple layers of Transformer encoders.

The self-attention mechanism is utilized to compute the degree of association between each word in the input sequence and the other words, generating weighted representations. Initially, given the input sequence $X = \{x_1, x_2, \ldots, x_n\}$, where n denotes the



"A man wears a salmon pink shirt with blue jeans and brown shoes. He carries a shopping bag. The person approached the source of fire."

"This person is wearing a pink long sleeved shirt and blue jeans. The are carrying a large white bag in left hand. Hair is black. Shoes appear to be brown. The person caught on fire."

"A man wears a pink shirt with blue jeans and brown shoes. He carries a shopping bag. The person approached the source of fire..."

The person approached the source of fire.

Prompt: High-risk scenes appear in the C5 area. Please leave the C5 area quickily.

Fig.3: *Our proposed early warning framework based on TIDCB model.*

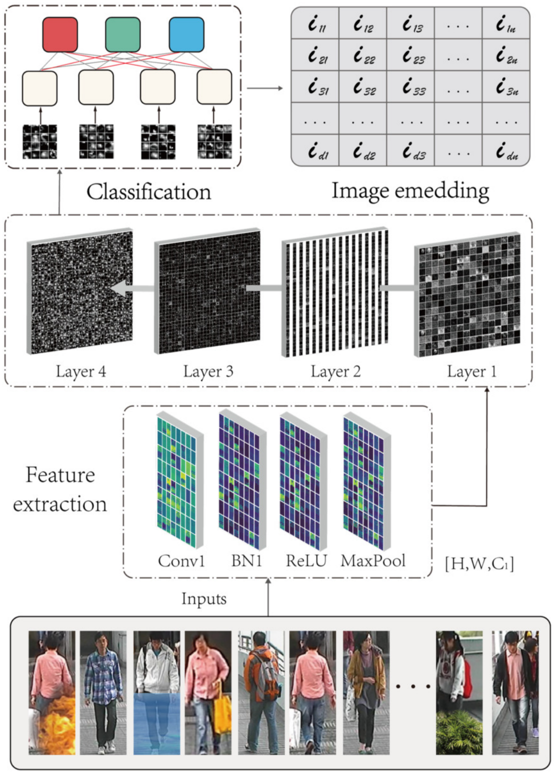


Fig.4: *The details of image feature extraction.*

sequence length and each x_i represents a word vector, three new sequences, $Q = XW_Q$, $K = XW_K$ and $V = XW_V$, are obtained through linear transformations, where W_Q, W_K and W_V are weight matrices. Then, the similarity score matrix $S = QK^T$ is computed between Q, K, and V, followed by scaling the score matrix as $S = \frac{S}{\sqrt{d_k}}$, where d_k is the dimensionality of the keys. Next, the softmax function is applied to transform the similarity scores into the attention weight matrix $A = softmax(S)$. Finally, the weight matrices are multiplied with the value matrix and aggregated to obtain the self-attention output matrix.

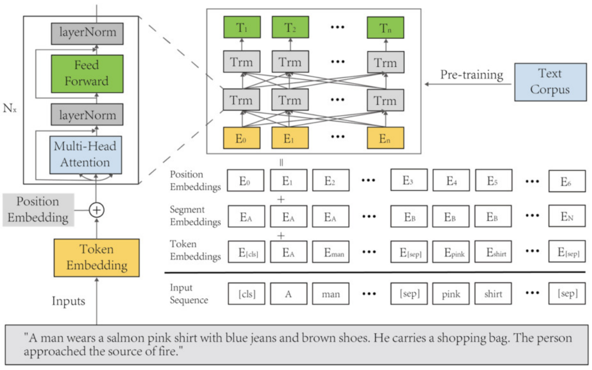
$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{1}$$

The feed-forward neural network performs nonlinear transformations and combinations of the self-attention outputs. In the Transformer, the feed-forward neural network comprises two fully connected layers interconnected by ReLU activation functions.

Given the self-attention output matrix Y, it is flattened into a vector $vec(Y)$. Then, through two linear transformations and the ReLU activation function, the intermediate representation is obtained as formula (2), where W_1 and W_2 are weight matrices, and b_1 and b_2 are bias vectors.

$$FFN(vec(Y)) = RELU(YW_1 + b_1)W_2 + b_2 \tag{2}$$

Finally, the intermediate representation is reshaped to match the shape of the self-attention output matrix, yielding the output matrix of the feed-forward neural network.



"A man wears a salmon pink shirt with blue jeans and brown shoes. He carries a shopping bag. The person approached the source of fire."

Fig.5: *The details of text feature extraction.*

3.2.3 Our loss function

The loss function used in the original model is Cross-Modal Projection Matching (CMPM)[24], which focuses on enhancing the image and text by minimizing the Kullback-Leibler (KL) dispersion between the projection compatibility distribution and the normalized matching distribution defined by all the positive and negative samples in the small batch matching between images and text. Specifically, the model first computes the normalized representations of the image and text embedding vectors and then performs a scalar projection to measure the similarity between them - i.e., the more similar the image features are to the text features, the larger the value of their scalar projection. Next, the relationship between sample pairs is identified by creating a label mask, $M_{t,i}$ for each pair, where $M_{t,i} = 1$ indicates that samples t and i have the same label, while $M_{t,i} = 0$ indicates that they have different labels. In addition, the model normalizes the matching distribution by computing the cross-entropy of the true matching distribution. Finally, the CMPM loss between image-to-text and text-to-image is calculated and returned.

The image-to-text CMPM loss can be expressed as:

$$L_{i2t} = \sum\nolimits_i \sum\nolimits_t p(i,t) \log \frac{p(i,t)}{q(i,t)} \tag{3}$$

Where $p(i,t)$ denotes the probability that the image matches the text under true matching and $q(i,t)$ denotes the probability that the image matches the text under model prediction, both of which are computed by the softmax function.

Similarly, the text-to-image CMPM loss can be expressed as:

$$L_{t2i} = \sum\nolimits_t \sum\nolimits_i p(t,i) \log \frac{p(t,i)}{q(t,i)} \tag{4}$$

The total CMPM loss is then a summation of L_{i2t} and L_{t2i} and can be expressed as:

$$L_{cmpm} = L_{i2t} + l_{t2i} \tag{5}$$

In image-to-text embedding learning, the computation of matching loss usually involves two-way considerations: on the one hand, image-to-text matching loss emphasizes that the similarity between a matched text and its corresponding image should be significantly higher than the similarity between a mismatched text and that image; on the other hand, text-to-image matching loss ensures that the text associated with a particular image is prioritized over the irrelevant text in terms of order. The core goal is to minimize the cosine similarity between pairs of positive samples to achieve more accurate matching in the embedding space of images and texts.

To further optimize the model performance, we introduce Triplet loss[29] on top of CMPM loss, where the goal of the triplet loss is to ensure that the distance between an anchor sample and a positive sample is smaller than the distance between that anchor sample and a negative sample. Firstly, the distance $D_{positive}$ between the anchor a and the positive sample p in the embedding space is computed, as well as the distance $D_{negtive}$ between the anchor a and the negative sample n. The distance metric we use here is Euclidean distance.
This can be represented as follows:

$$D_{positive} = \sqrt{\sum\nolimits_{i=1}^{n} (a_i - p_i)^2} \tag{6}$$

$$D_{negative} = \sqrt{\sum\nolimits_{i=1}^{n} (a_i - n_i)^2} \tag{7}$$

Where a_i, p_i, and n_i represent the coordinates of the i-th dimension of the anchor, positive sample, and negative sample in the embedding space, respectively. Then, the triplet loss is computed and returned based on a given threshold value θ, which determines the minimum distance between the positive and negative samples. Proper selection of this parameter can further optimize the model's performance, as demonstrated in the experimental section of Chapter 4. Finally, the triplet loss is defined as:

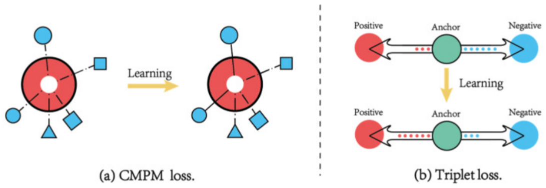
$$L_{triplet} = \max(D_{positive} - D_{negative} + \theta, \quad 0) \tag{8}$$

The objective of the loss function is to minimize the distance between samples of the same class while pushing apart samples from different classes. If $D_{positive} - D_{negative} + \theta > 0$, the triplet samples are correctly arranged in the embedding space. Otherwise, a non-zero loss value signifies that the arrangement of triplet samples is incorrect and needs adjustment through methods such as gradient descent to minimize the loss function. Adding the Triplet loss introduces new learning constraints to the model, facilitating the proximity of samples from the same class in the embedding space while distancing samples from different classes.

Thus, the total loss $L_{CMPM-Triplet}$ is the sum of the CMPM and triplet loss:

$$L_{CMPM-Triplet} = L_{cmpm} + L_{triplet} \tag{9}$$

Overall, the CMPM-Triplet loss function comprehensively considers the cross-modal matching relationship between images and text. The CMPM loss focuses on enhancing the matching between images and text by minimizing the KL divergence to optimize the feature projection matching between the two modalities. Meanwhile, the Triplet loss introduces cross-sample triplet relationships to ensure that samples of the same category are closer to each other in the embedding space while samples of different categories are farther apart, thereby enhancing the accuracy and robustness of cross-modal matching, as shown in Fig. 6 . The CMPM-Triplet loss function enables the model better to understand the complex relationship between images and text, thereby improving the effectiveness of matching tasks. Moreover, it exhibits strong generalization capabilities when handling different datasets and tasks, which will be further validated in our experimental section.



(a) CMPM loss.

(b) Triplet loss.

Fig.6: *CMPM loss and Triplet loss.*

3.3 New Early Warning framework based on TIDCB model

This early warning framework can be applied in communities, streets, and shopping malls, as demonstrated in Fig. 7. Any space can be artificially divided into several areas, and similarly, in this study, we have divided the floor space of a shopping mall into zones ranging from A1 to E5, which could correspond to real-world divisions such as clothing, food, and electronics areas. In each region, several cameras are deployed with predetermined positions and assigned numbers, as shown in Fig. 8. For example, when a fire breaks out, our TIDCB model analyzes the features of individuals and the surrounding environment captured in the surveillance footage, generating textual descriptions and assessing the predetermined risk levels. In this experiment, considering the complexity of social behavior, we initially set “fire” as a high-risk scenario, “water” as a medium-risk scenario, and “grass” as a low-risk scenario. Upon detecting a potential security incident in the mall, such as “fire” from our TIDCB model, cameras B2, B3, C2, and C3 quickly pinpointed C5 and its surrounding area, providing feedback to our alert framework.

In this instance, “a man” and “fire” are the learned character and scene traits; more precisely, “A young man with short hair wearing a black and white striped sweater, black trousers and black-rimmed glasses is very close to the source of the fire.” is the generated text description. As a result, the warning system in the mall broadcasts the following message: “High-risk scenes appear in the C5 area. Please leave the C5 area quickly.” The warning device indicates that individuals in each region respond swiftly, leaving the C5 area and heading for the closest safe passage, as shown in Figure (c). For instance, the warning “Please move away from the C5 area and evacuate towards gate 1” would have been delivered to those in the B2 area. Allowing individuals to flee without entering the danger zone should reduce or prevent casualties.

4. EXPERIMENTS

4.1 Experimental setup

Datasets. We created a custom dataset called CUHK-PEDES-DANGER in our work. It comprises three parts: training, testing, and validation sets. There are a total of 2,385 pedestrian examples, 19,080 images, and 38,160 text summaries in the training and testing sets. Setting the “test_size” option to 0.076 randomly splits the data during training. This meant that 7.6% of the whole dataset was used for the testing set, and 92.4% was used for the training set. The validation set contains 480 pedestrian instances, different people from the training and testing sets, 1,920 images, and 3,840 text descriptions, distinct from the data in the training and testing sets. Each dataset has images and text explanations that go with them. All the photos were processed and resized to the same size of 384x128 pixels so that the data would be consistent.

Implementation detail. This study conducted model training on an RTX 3080 Ti GPU. We utilized the PyTorch framework, employing torch 1.10.2, torch_vision 0.11.3, Python 3.11, and CUDA 11.3. The training process involves several key parameters and practices, and we use Adam Optimizer, with a learning rate of 0.003, reduced by 0.1 times after 50 cycles. Each training consists of 64 image-text pairs to ensure efficient use of computing resources. The model trained a total of 100 cycles, of which the initial 10 cycles were in the heating phase. Weight reduction is set to, and as needed, use multiple learning rate reduction strategies, such as “Multi-step LR,” “Step LR” or “Low LR on Plateau.” In addition, we have an early stop mechanism that stops training if val_rank1 is not improved for 10 consecutive cycles. Model checkpoints are kept regularly to capture optimal performance based on verified accuracy. In order to compare our method with others, we have further selected TIPCB as the baseline model and compared our method with their dataset.

4.2 Experiment of θ parameters

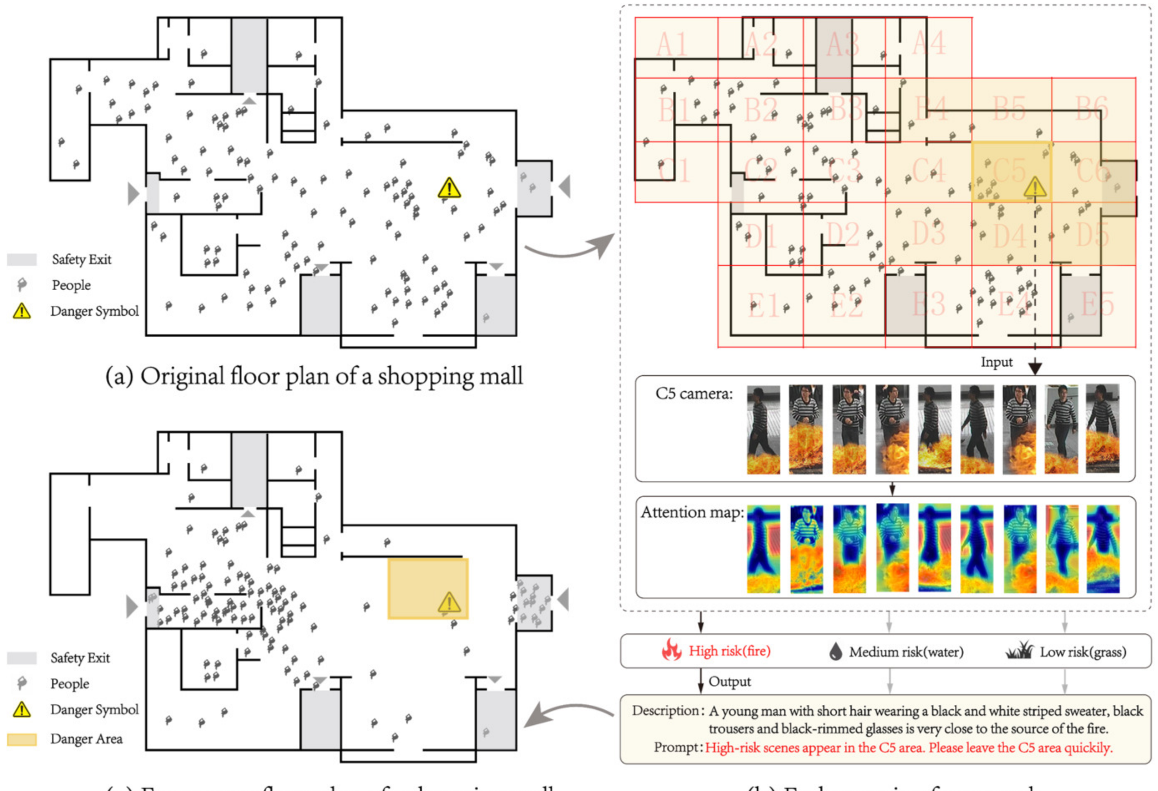
Because the θ in the CMPM-Triplet loss function will have a certain impact on the experimental accuracy, it determines the minimum value of the distance between the positive and negative samples. So we have experimented with different θ values, and from the experimental results in Table 2, it seems that different θ values do not affect the rank value too much, so we took the middle value of 0.5 to carry out the later experiments.

Table 2: *The effect of different θ values in our model on experimental accuracy.*

θ	Text to Image Rerank				Image to Text Rerank			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
0.1	0.69218	0.87165	0.92965	0.48026	0.69389	0.90803	0.95539	**0.44485**
0.4	0.70110	0.87440	0.92519	**0.48891**	**0.70076**	**0.92519**	**0.96568**	0.44160
0.5	**0.70590**	**0.87509**	0.92656	0.48874	0.69732	0.92107	0.96156	0.43691
0.6	0.69286	0.87234	0.92965	0.48347	0.68909	0.92176	0.96156	0.43333
0.9	0.69218	0.87303	**0.93342**	0.48551	0.69046	0.92107	0.96362	0.44174

Table 3: *Comparison of the CMPM loss with CMPM-Triplet loss(our loss) on the CUHK-PEDES-DANGER dataset.*

Loss	Batch size	Text to Image Rerank				Image to Text Rerank			
		Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
CMPM	16	0.69357	0.85463	0.91891	0.40782	0.75112	0.91256	0.95217	0.38688
Ours	16	0.69389	0.87440	0.92896	0.49615	0.71723	0.92382	0.96019	0.45123
CMPM	64	**0.73729**	0.87108	**0.92975**	0.48866	**0.75187**	0.92676	0.96413	0.44818
Ours	64	0.73430	**0.87818**	0.92377	**0.54840**	0.73916	**0.93199**	**0.96786**	**0.49812**



(a) Original floor plan of a shopping mall

Description: A young man with short hair wearing a black and white striped sweater, black trousers and black-rimmed glasses is very close to the source of the fire.

Prompt: High-risk scenes appear in the C5 area. Please leave the C5 area quickly.

(c) Emergency floor plan of a shopping mall

(b) Early warning framework

Fig.7: *Demonstration of our warning scheme in a shopping mall based on TIDCB model.*

4.3 Comparison of loss functions

We conducted extensive experiments using the CUHK-PEDES-DANGER dataset on both the CMPM and CMPM-Triplet loss functions. In these experiments, we particularly focused on the impact of different batch sizes (16 and 64) on the performance of “text-to-image” and “image-to-text” retrieval tasks, especially when using re-ranking post-processing.

To validate the effectiveness of the CMPM-Triplet loss, we conducted comparative experiments to assess the performance of the CMPM-Triplet loss function against the original CMPM loss function. In our experiments, we specifically examined the impact of different batch sizes (16 and 64) on the performance of “Text-to-Image” and “Image-to-Text” retrieval tasks, particularly when using post-processing with re-ranking. The training accuracy of both loss functions on the CUHK-PEDES-DANGER dataset is illustrated in Fig. 9, showing that after 20 training epochs, our loss function surpassed CMPM in accu-

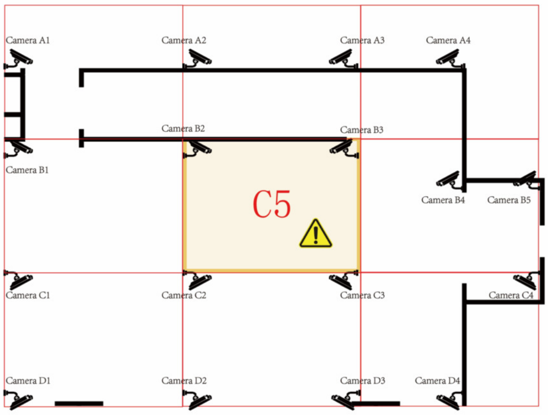


Fig.8: *Demonstration diagram of camera distribution in a shopping mall.*

racy and continued to improve, consistently outperforming CMPM after that steadily. Table 3 highlights the higher accuracy and scores of our loss function at a batch size of 64. This demonstrates that our method exhibits stable and excellent performance across various evaluation metrics. Moreover, our method showed good robustness and flexibility when adjusting for different batch sizes, which is crucial for adaptability in real-world application scenarios.

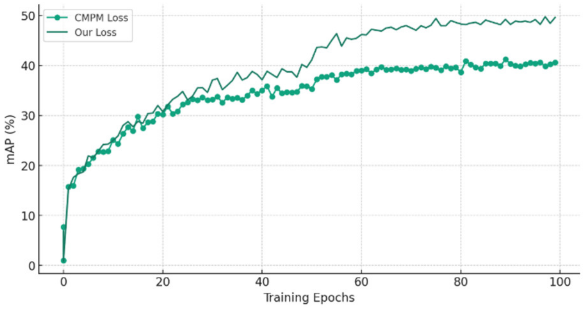


Fig.9: *Comparison of mAP values between CMPM Loss and CMPM-Triplet loss on the CUHK-PEDES-DANGER dataset.*

Table 4: *Comparison of the TIPCB model(baseline) with TIDCBour model) on the CUHK-PEDES dataset.*

Method	TIPCB (baseline)	Ours
Model	ResNet50+BERT Structure	ResNet50+RoBERTa Structure
Loss	CMPM	CMPM+Triplet
Rank-1	63.63	73.43(↑9.80)
Rank-5	82.82	87.82(↑5.00)
Rank-10	89.01	92.38(↑3.37)

4.4 Comparison of TIPCB and TIDCB

We conducted extensive experiments using the CUHK-PEDES dataset on TIPCB and our proposed model (TIDCB). The improvement of our model over TIPCB lies in replacing BERT with RoBERTa and introducing a novel loss function, CMPM-Triplet. Table 4 presents the performance comparison of the two methods on the original CUHK-PEDES dataset. It can be observed that our model achieves higher scores in Rank-1, Rank-5, and Rank-10.

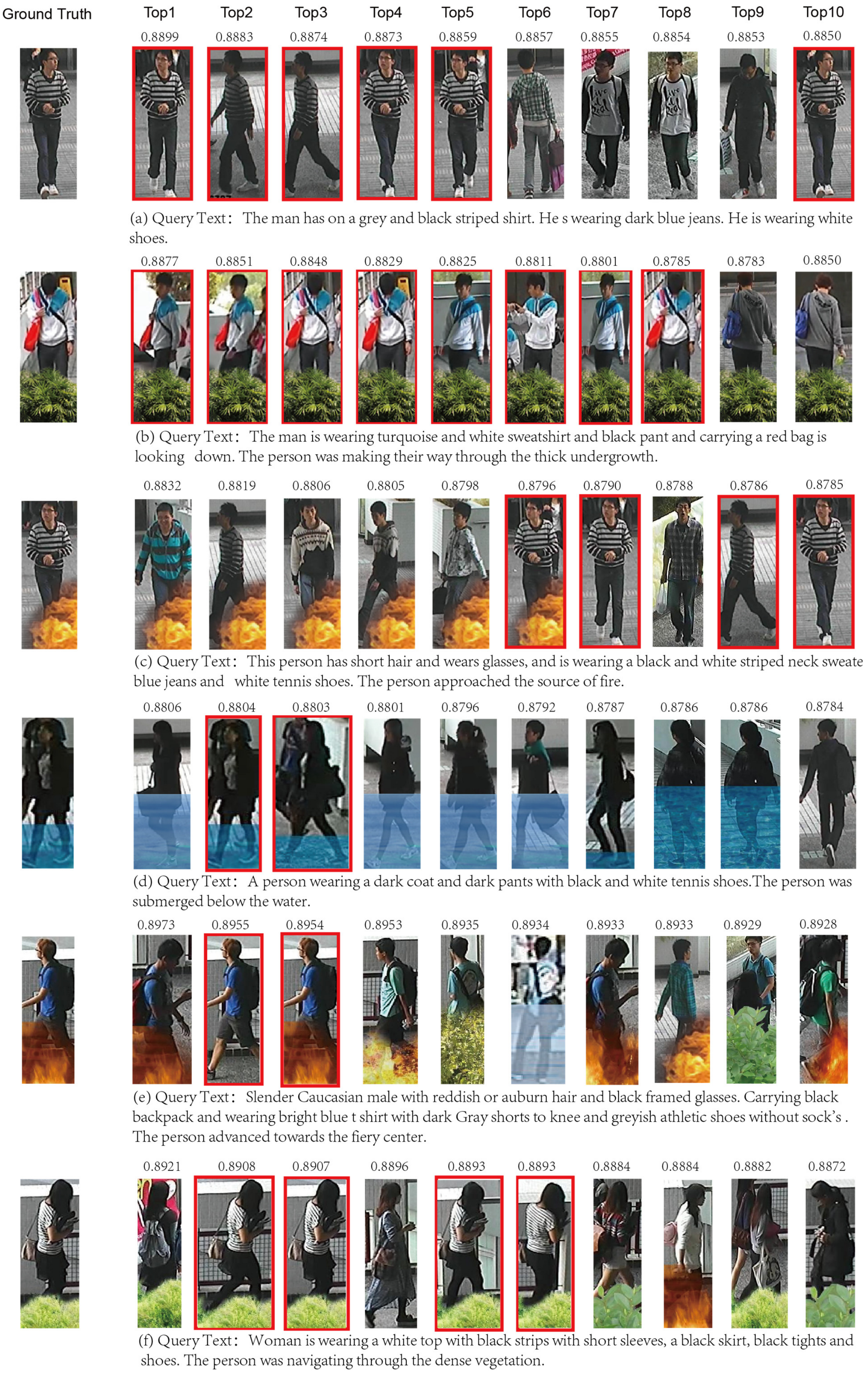
4.5 Actual results on the CUHK-PEDES-DANGER dataset

Table 5 shows how well the two models did on validation datasets in a comparison test. We chose examples from 52 sets of text and picture pairs to look at. The matching rate for the TIPCB model was 71.15%, and the matching rate for our model was 76.93%. Our model made it by 4.78 percentage points higher than the previous model. Our models are better at dealing with complicated situations because they match 18 (34.62%) of the time, 19 (36.54% of the time), and 3 (5.77% of the time) in the top 1, top 5, and top 10 levels of matching. When there was a non-match, our model had only 12 (23.08%) errors, while the TIPCB model had 15 (28.85%) errors.

The results of validating our method are shown in

Table 5: *Actual results on the CUHK-PEDES-DANGER dataset.*

Method	Match Results		Number of Matches (Match Rate)	Cumulative Number of Matches (Cumulative Match Rate)
TIPCB	Match	top1	22（42.31%）	22（42.31%）
		top5	10（19.23%）	32（61.54%）
		top10	5（9.61%）	37（71.15%）
	Mismatch		15（28.85%）	15（28.85%）
Ours	Match	top1	18（34.62%）	18（34.62%）
		top5	19（36.54%）	37（71.16%）
		top10	3（5.77%）	**40（76.93%）**
	Mismatch		12（23.08%）	**12（23.08%）**



(a) Query Text： The man has on a grey and black striped shirt. He s wearing dark blue jeans. He is wearing white shoes.

(b) Query Text： The man is wearing turquoise and white sweatshirt and black pant and carrying a red bag is looking down. The person was making their way through the thick undergrowth.

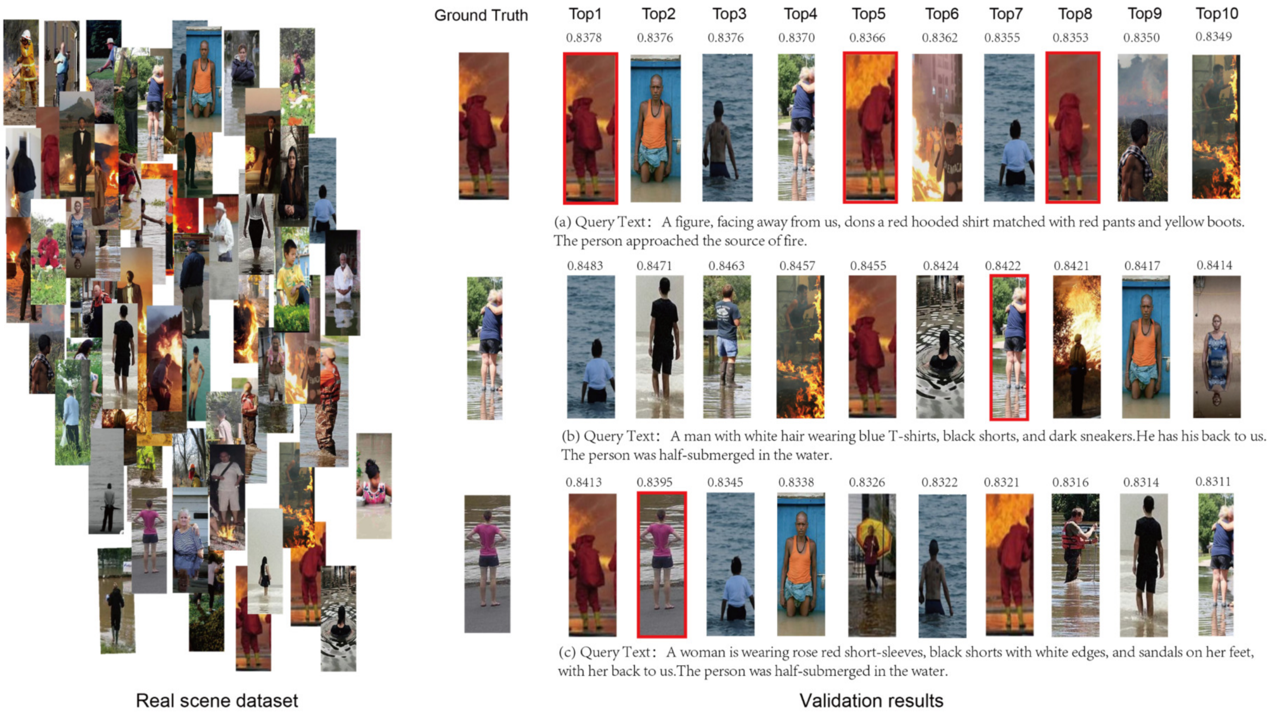
(c) Query Text： This person has short hair and wears glasses, and is wearing a black and white striped neck sweate blue jeans and white tennis shoes. The person approached the source of fire.

(d) Query Text： A person wearing a dark coat and dark pants with black and white tennis shoes.The person was submerged below the water.

(e) Query Text： Slender Caucasian male with reddish or auburn hair and black framed glasses. Carrying black backpack and wearing bright blue t shirt with dark Gray shorts to knee and greyish athletic shoes without sock's . The person advanced towards the fiery center.

(f) Query Text： Woman is wearing a white top with black strips with short sleeves, a black skirt, black tights and shoes. The person was navigating through the dense vegetation.

Fig.10: *Typical retrieval results. The right match is the image with the red box. The number at the top of each picture shows how similar it is thought to be to a given text query.*

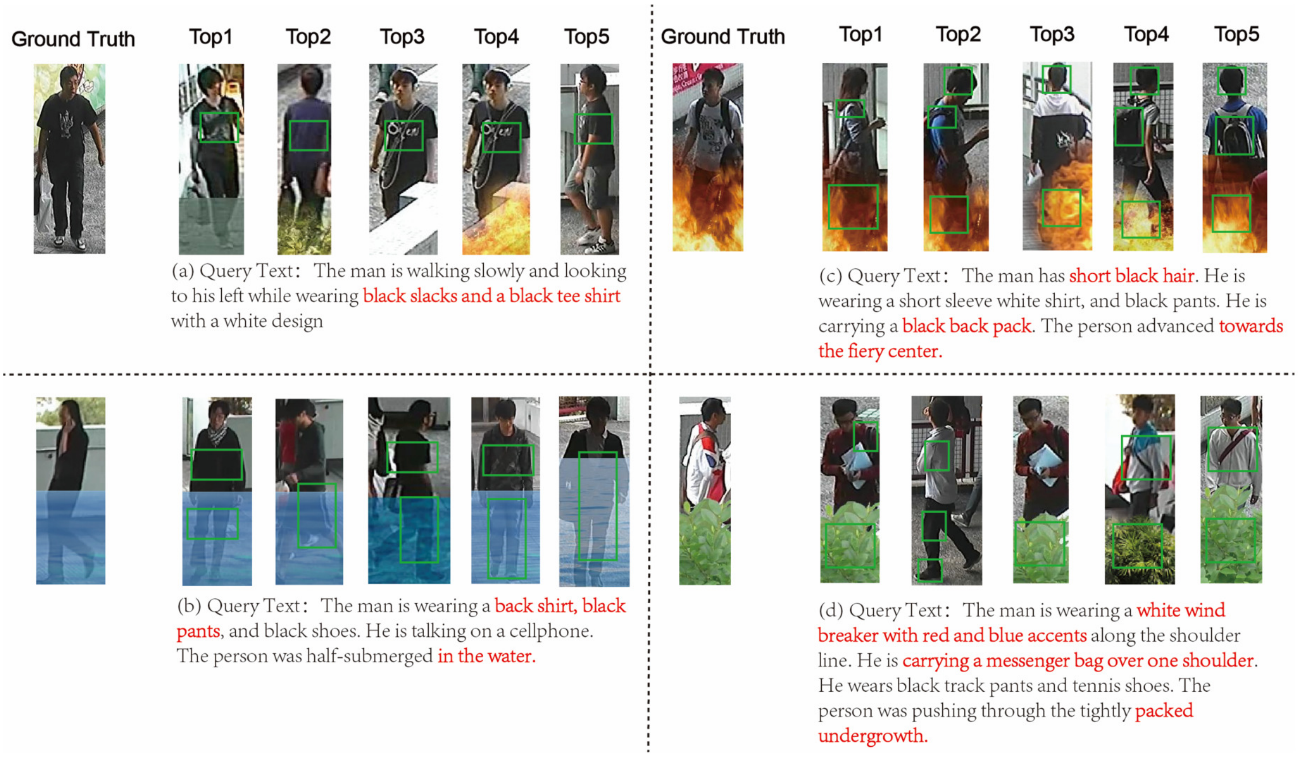


(a) Query Text： A figure, facing away from us, dons a red hooded shirt matched with red pants and yellow boots. The person approached the source of fire.

(b) Query Text： A man with white hair wearing blue T-shirts, black shorts, and dark sneakers.He has his back to us. The person was half-submerged in the water.

(c) Query Text： A woman is wearing rose red short-sleeves, black shorts with white edges, and sandals on her feet, with her back to us.The person was half-submerged in the water.

Fig.11: *Verification results in real scenarios using TIDCB model.*



(a) Query Text： The man is walking slowly and looking to his left while wearing black slacks and a black tee shirt with a white design

(c) Query Text： The man has short black hair. He is wearing a short sleeve white shirt, and black pants. He is carrying a black back pack. The person advanced towards the fiery center.

(b) Query Text： The man is wearing a back shirt, black pants, and black shoes. He is talking on a cellphone. The person was half-submerged in the water.

(d) Query Text： The man is wearing a white wind breaker with red and blue accents along the shoulder line. He is carrying a messenger bag over one shoulder. He wears black track pants and tennis shoes. The person was pushing through the tightly packed undergrowth.

Fig.12: *Failed retrieval results. The green box represents some of the correct features identified.*

Fig. 10. Putting in a specific text query will show you the top 10 image retrieval results. The pictures with the red boxes show that the text matches correctly. The numbers above each picture show the matching scores, which show how similar the picture is to the text description. Based on the results, it is clear that our models can successfully retrieve the image of a particular scene when text descriptions include specific scene features. Even though the results 7, 8, 9, and 10 don't show the scene elements that were directly asked for (like "fire source"), the characters in them match what was described in the text queries: "short," "brown glasses," "black and white clothes," "blue jeans," and "white net boots." This proof further shows how well our model works even when it is not in a specific scene and points to a very good account of what happened.

Because our dataset comprises artificially generated images of hazardous scenarios using artificial intelligence techniques, such hazardous scenes are added to a real image, lacking real-world hazardous situations. To test the possibility of applying our method to real-world hazardous situations, we gathered 50 images of real hazardous situations from the internet as a validation set due to the scarcity of hazardous images in real-world scenarios. The data we gathered from the real world differs greatly from our dataset. First, the hazardous scene differs from our data; second, there is only one image per person, compared to 8 images per person from different angles in our dataset. We tested our model trained in our dataset and applied it to this real-world data set. As shown in Fig. 11, it can be observed that our model is capable of detecting some hazardous situations in the real world with a low accuracy, top 10 at 34%. This is primarily attributed to the diversity of our dataset. In the future, we aim to expand our dataset and enhance its diversity to improve the accuracy of the model.

Certainly, our model also exhibits certain limitations, as illustrated in Fig. 12, which displays some failure cases encountered during the validation process. In these instances, the green boxes denote the identified features. For instance, in the first group (a), the query text describes a man walking slowly, looking left, wearing black trousers and a black T-shirt with white patterns. However, the images ranging from "Top1" to "Top5" depict males wearing similar attire, yet with varying postures and backgrounds, failing to match the "real image completely."

This is because our model heavily relies on the quality and comprehensiveness of textual descriptions; if the text descriptions are ambiguous or incomplete, our model's performance may degrade. Secondly, the diversity in human appearance, attire, and behavior can lead to a wide range of potential images, making it challenging for the model to find precise matches. Lastly, the model may not have been trained on a sufficiently diverse dataset, lacking depth in understanding complex scenes. These limitations will be addressed in future studies further to improve the performance and applicability of our model.

5. CONCLUSION

This study proposes a text-based warning framework. Previous research on text-based descriptions has primarily focused on pedestrian search by matching images of pedestrian body parts to text, lacking the search for pedestrians in dangerous scenarios. This study aims to achieve rapid retrieval of individuals in hazardous situations to prevent or effectively respond to crises by matching environmental and pedestrian body part images to text. Through improvements and expansions on the CUHK-PEDES dataset, we constructed the CHUK-PEDES-DANGER dataset, which includes various hazardous scenarios, providing valuable resources for research in this field. Our TIDCB model builds upon the TIPCB model by integrating ResNet-50 and RoBERTa models to extract image and text features and introduces a novel CMPM-Triplet loss function to achieve effective matching of cross-modal features. On the validation dataset, our model outperforms TIPCB with a matching rate of 76.93%, an improvement of 4.78% compared to TIPCB, and demonstrates significant advantages in handling complex scenarios.

In the future, we can add more types of images of dangerous scenes to the dataset based on different use cases so that it can handle a wider range of hazardous scenes. At the same time, our system is flexible enough to work with different languages. It is currently focused on English, but its structure means it can be changed to work with other languages in the future if needed.

Finally, we will look at how the framework can be connected to video security and emergency response systems that are already in place to help make communities smarter and safer.

The source code and dataset for this study can be found at `https://github.com/Zfofo/TIDCB`.

ACKNOWLEDGEMENT

The first author is awarded with CPALL scholarship while conducting this research in PIM.

The first and last authors each contributed 50% equally to this work. The last author is the corresponding author.

AUTHOR CONTRIBUTIONS

Conceptualization, FF.Z. and J. Q.; methodology, FF.Z. and J. Q.; software, FF.Z and J. Q.; validation, FF.Z. and J. Q.; formal analysis, FF.Z. and J. Q.; investigation, FF.Z. and J. Q.; data curation, FF.Z. and J. Q.; writing—original draft preparation, FF.Z.

and J. Q.; writing—review and editing, FF.Z. and J. Q.; visualization, FF.Z. and J. Q.; supervision, J. Q.; All authors have read and agreed to the published version of the manuscript.

References

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao and S. C. Hoi, "Deep learning for person reidentification: A survey and outlook," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no.6, pp. 2872-2893, 2021.

[2] Y. Chen, G. Zhang, Y. Lu, Z. Wang and Y. Zheng, "Tipcb: A simple but effective partbased convolutional baseline for text-based person search," *Neurocomputing*, vol. 494, pp. 171-181, 2022.

[3] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue and X. Wang, "Person search with natural language description," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1970-1979, 2017.

[4] X. Han, S. He, L. Zhang and T. Xiang, "Text-based person search with limited data," *arXiv preprint*, 2021. [Online]. Available:https://doi.org/10.48550/arXiv.2110.10807.

[5] S. Choi and S. Yi,(2022,Septemper 27). "Seven Killed and One Seriously Injured in a Fire at Hyundai Outlet Daejeon Store: Decision Expected on the Application of the Serious Accidents Punishment Act," https://www.khan.co.kr/national/incident/article/202209261810001.

[6] W. Ngamkham,(2023,Octorber 3). "Two dead in shopping mall shooting," https://www.bangkokpost.com/thailand/general/2656821/two-dead-in-shopping-mall-shooting.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," arXiv preprint,2019. [Online]. Available: https://doi.org/10.48550/arXiv.1907.11692.

[8] H. Yang, P. Liu, S. Li,H. Liu and H. Wang, "A real-time framework for dangerous behavior detection based on deep learning," *Proceedings of the 2022 4th International Conference on Robotics, Intelligent Control and Artificial Intelligence*, pp. 1200-1206, 2022.

[9] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," *2017 IEEE international conference on image processing (ICIP)*, pp. 1577-1581, 2017.

[10] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee and A. Y. Ng, "Multimodal deep learning," *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 689-696, 2011.

[11] N. Srivastava and R. Salakhutdinov, "Learning representations for multimodal data with deep belief nets," *International conference on machine learning workshop*, pp. 978-1, 2012.

[12] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 207-218, 2014.

[13] F. Feng, X. Wang and R. Li, "Cross-modal retrieval with correspondence autoencoder," *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 7-16, 2014.

[14] W. Wang, B. C. Ooi, X. Yang, D. Zhang and Y. Zhuang, "Effective multi-modal retrieval based on stacked auto-encoders," *Proceedings of the VLDB Endowment*, vol. 7, no.8, pp. 649-660, 2014.

[15] D. Wang, P. Cui, M. Ou and W. Zhu, "Learning compact hash codes for multimodal representations using orthogonal deep structure," *IEEE Transactions on Multimedia*, vol. 17, no.9, pp. 1404-1416, 2015.

[16] H. Zhang, Y. Yang, H. Luan, S. Yang and T.-S. Chua, "Start from scratch: Towards automatically identifying, modeling, and naming visual attributes," *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 187-196, 2014.

[17] H. Zhang, X. Shang, H. Luan, M. Wang and T.-S. Chua, "Learning from collective intelligence: Feature learning using social images and tags," *ACM transactions on multimedia computing, communications, and applications (TOMM)*, vol. 13, no.1, pp. 1-23, 2016.

[18] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu and S.Yan, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE transactions on cybernetics*, vol. 47, no.2, pp. 449-460, 2016.

[19] Y. Peng, X. Huang and J. Qi, "Cross-media shared representation by hierarchical learning with multiple deep networks," *IJCAI*, pp. 3853, 2016.

[20] Y. Peng, J. Qi, X. Huang and Y. Yuan, "CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Transactions on Multimedia*, vol. 20, no.2, pp. 405-420, 2017.

[21] J. Qu and A. Shimazu, "Cross-language information extraction and auto evaluation for OOV term translations," *China Communications*, vol. 13, no.12, pp. 277-296, 2016.

[22] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu and Y.-D. Shen, "Dual-path convolutional

image-text embeddings with instance loss," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no.2, pp. 1-23, 2020.

[23] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," *Proceedings of the European conference on computer vision (ECCV)*, pp. 686-701, 2018.

[24] N. Sarafianos, X. Xu and I. A. Kakadiaris, "Adversarial representation learning for text-to-image matching," *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5814-5824, 2019.

[25] Y. Chen, R. Huang, H. Chang, C. Tan, T. Xue and B. Ma, "Cross-modal knowledge adaptation for language-based person search," *IEEE Transactions on Image Processing*, vol. 30, pp. 4057-4069, 2021.

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

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[28] Q. Li and J. Qu, "A novel BNB-NO-BK method for detecting fraudulent crowdfunding projects," *Songklanakarin Journal of Science & Technology*, vol. 44, no.5, 2022.

[29] A. Hermans, L. Beyer and B. Leibe, "In defense of the triplet loss for person reidentification," *arXiv preprint*, 2017. [Online]. Available: `https://doi.org/10.48550/arXiv.1703.07737`.

Fangfang Zheng is currently studying for the Master of Engineering Technology, Faculty of Engineering and Technology, Panyapiwat Institute of Management, Thailand. She received B. Arch from Nanjing Tech University Pujiang Institute, China, in 2022. Her research interests are artificial intelligence, natural language processing, image processing and information retrieval.

Jian Qu is an Assistant Professor at the Faculty of Engineering and Technology, Panyapiwat Institute of Management. He received Ph.D. with Outstanding 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 severing as a house committee for Thai SUPERAI project since 2020. His research interests are natural language processing, intelligent algorithms, machine learning, machine translation, information retrieval, image processing, and autonomous driving.