scientific reports

Check for updates

OPEN

Enhanced multi-scale trademark element detection using the improved DETR

Longwen Li[1], Xiuhui Wang[1]✉ & Wei Qi Yan[2]✉

The exponential growth in the number of registered trademarks, coupled with the escalating incidents of trademark infringement, has made the automatic detection of such infractions a crucial area of study in the domain of market regulation. In light of the diverse range of elements and the pervasive presence of small targets in trademark images, we present an enhanced version of the DETR-based Multi-Scale Trademark Element Detection Network (MSTED-Net). Our primary innovation lies in incorporating a dual fusion mechanism that integrates the Spatial Attention Module (SAM) and Global Context Network (GCNet) within the backbone network, thereby providing a more robust approach to capture the essential characteristics of the trademark images under investigation. Subsequently, we develop a Multi-scale Feature Augmentation Pyramid (MFA-FPN), which aims to further fortify the model's ability to extract features and boost the detection efficiency for small targets. The efficacy of our proposed detection network is demonstrated through experimental results, showcasing an outstanding detection accuracy of 91.12% in comparison to other state-of-the-art detection algorithms.

Keywords Detection transformer, Attention mechanism, Multi-scale feature fusion, Trademark retrieval

Trademarks serve as a vital form of intellectual property that not only embody the image of a brand but also bolster its competitive edge in today's densely packed market landscape. With the swift progression of internet technology and the emergence of e-commerce platforms, trademark databases have grown exponentially larger. The overwhelming volume of trademark image data has intensified the challenges associated with protecting this type of intellectual property, leading to numerous instances of trademark infringement. By considering the immense scope of trademark image databases, it has become impractical to rely solely on human labor for detecting identical or confusingly similar trademarks in the context of infringement assessments.

The complexity of trademark detection is primarily manifested in several aspects. Firstly, trademarks can appear against complex backgrounds such as billboards, packaging boxes, product displays, etc., necessitating not only accurate localization of the trademarks but also the exclusion of interference from distracting objects. Secondly, there exists diversity in the form and style of trademarks; some are purely textual while others incorporate complex graphic elements. This variety demands a higher level of robustness from the detection model. Furthermore, the scale and orientation of trademark occurrences are also highly variable, particularly when dealing with small-scale trademarks where the challenges become more pronounced. In the context of trademark detection, small-scale trademarks refer to elements that occupy an extremely small proportion within an image, typically ranging from 2% to 5% of the entire image size. Despite their minute presence, these small-scale trademarks often exhibit intricate details, especially when they consist of combinations of text, patterns, and symbols, which augment the complexity of detection. Additionally, due to the low pixel occupancy of small-scale trademarks within images, noise or other visual elements in the background may significantly interfere with them, making it difficult for models to effectively segregate the trademarks from the surrounding environment.

In light of the rapid advancements in deep learning, constructing a multisource trademark infringement recognition system has become indispensable[1]. As depicted in Fig. 1, this system classifies trademarks into three categories: graphic trademarks, mixed graphic-text trademarks, and text trademarks. Focusing on the analysis, extraction, and information extraction of graphic trademarks, our research aims to provide a more precise determination of potentially infringing trademark images and their similarities through the identification of trademark elements and attribute analysis. Improvements in trademark detection algorithms have been achieved by incorporating artificial feature operators[2–4], which have resulted in subpar detection accuracy, complicated computational processes, and restricted generalizability. The advent of convolutional neural networks[5–7],

[1]China Jiliang University, Hangzhou 310018, China. [2]Auckland University of Technology, Auckland, New Zealand. ✉email: wangxiuhui@cjlu.edu.cn; weiqi.yan@aut.ac.nz

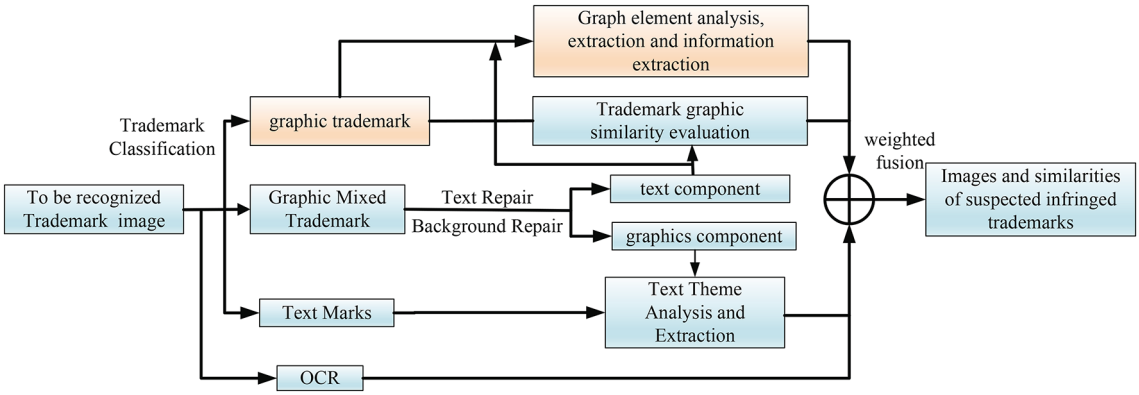


Figure 1. Multi-source trademark infringement detection.

however, has propelled the use of deep learning for trademark detection[8–10], leading to notable performances that far surpass traditional algorithms based on artificial feature operators. These technological advancements have created new opportunities for trademark detection tasks and offer sturdy technical support for identifying and combating counterfeit trademarks.

This paper presents an enhanced version of the DETR-based trademark element detection network, referred to as the Multi-Scale Trademark Element Detection Network (MSTED-Net). In comparison to other state-of-the-art approaches, MSTED-Net exhibits outstanding performance. The primary contributions of this paper are grouped into three important aspects:

- **For the first time, the DETR framework was advanced and implemented for trademark element detection, leading to the development of a multi-scale trademark element detection network called MSTED-Net.** Optimizing MSTED-Net resulted in a simple yet efficient model that effectively captures more feature information about small objects, empowering the model to focus more intently on small object detection within trademark images.
- **A novel attention mechanism, referred to as the Dual Perception Attention Module (DPAM), was proposed**. DPAM fuses global and spatial information through a unique dual fusion approach, enabling the model to adapt to complex scenes with varying target sizes in trademark images and accurately locate trademark element regions. This considerable improvement significantly enhances the overall detection performance of the model.
- **A Multi-scale Feature Augmentation Pyramid (MFA-FPN) was designed to effectively integrate features of trademark images at different scales.** By providing abundant feature information for trademark element detection, it achieves an effective fusion of high-level semantic and low-level detailed information, guaranteeing comprehensive application of cross-scale features and further improving detection performance. The fusion of shallow and deep features ensures even greater effectiveness in detecting trademark elements.

Related work

In the domain of object detection, the task of trademark identification has been significantly advanced by leveraging deep learning algorithms. This area is characterized by its classification into two predominant methodological streams: two-stage approach and one-stage technique[11].

Two-stage algorithms predominantly utilize Convolutional Neural Networks (CNNs), with the pioneering framework being the Regional-based Convolutional Neural Networks (R-CNN)[12–14]. The R-CNN architecture initiates by employing Selective Search to produce candidate regions, followed by the utilization of a CNN for both classification tasks and bounding box localization. Nonetheless, the requirement to isolate and extract features for each individual candidate region incurs substantial computational overhead, leading to a notably slower inference rate. In response to this challenge, the Spatial Pyramid Pooling networks (SPP-nets)[15–17] were introduced. These networks facilitate the extraction of shared features across the entire image, employ a spatial pyramid pooling layer for aggregation, and rely on Support Vector Machines (SVMs) and linear regressors for subsequent classification and regression operations. While SPP-nets substantially alleviated the inference bottleneck inherent in R-CNN, they concurrently augmented both training duration and algorithmic complexity. To streamline the integration of feature extraction and classification, Fast R-CNN[18,19] evolved as an enhancement over R-CNN, consolidating these processes within a unified network. By extracting feature maps from the full image via a CNN, implementing region-of-interest (ROI) pooling, and subsequently performing classification and bounding box regression, Fast R-CNN achieved a notable improvement in both detection speed and overall network efficiency through comprehensive end-to-end training. Building further upon this foundation, Faster R-CNN[20–22] refined the candidate region proposal mechanism by incorporating a Region Proposal Network (RPN). This RPN operates directly on the convolutional feature map to propose candidate regions, utilizes a sliding window strategy to create anchor points, and conducts binary classification (distinguishing between foreground and background) along with bounding box regression. This innovation streamlined the region proposal procedure and dramatically expedited the detection pace. Despite the commendable performance of

two-stage detection algorithms, their intricate architectural design necessitates multi-step procedures for feature acquisition and candidate box determination. Consequently, they exhibit elevated computational demands and detection latency, posing challenges when applied to practical scenarios involving real-world trademark imagery.

Advancements in the field have culminated in the development of one-stage object detection algorithms, which offer superior computational efficiency. Characterized by their integrated approach, these algorithms converge the classification and detection functions. They operate by densely and uniformly sampling objects across diverse positions, scales, and aspect ratios during the detection phase. Notably, the prominent one-stage object detection algorithms are Single Shot Multibox Detector (SSD)[23–25], You Only Look Once (YOLO)[26–30], and RetinaNet[31–33]. SSD exemplifies its category by detecting objects directly on feature maps of varying scales. It deploys multiple predefined bounding boxes per feature map and executes classification and bounding box regression on these candidates. The principal strength of SSD lies in its efficiency; it achieves rapid detection rates without compromising accuracy, particularly for larger and mid-size objects. However, it struggles with smaller objects, exhibiting suboptimal performance. RetinaNet represents another milestone that markedly enhances detection capabilities for challenging targets through the introduction of Focal Loss. Despite this advancement, RetinaNet retains a relatively higher computational cost compared to other models. The YOLO family of algorithms has revolutionized traditional detection tasks by streamlining the detection process and refining the underlying framework. Their exceptional performance in numerous applications is well documented. Nevertheless, YOLO inherently grapples with issues such as extensive parameter sizes and computational requirements, which limit its deployment in resource-restricted settings. In summary, one-stage object detection algorithms present compelling benefits in terms of detection speed and architectural simplicity. Yet, they continue to encounter challenges pertaining to detection precision under certain specific conditions, where fine-grained or subtle object characteristics pose difficulties. Similar approaches utilizing multi-scale pyramids were also applied in Umirzakova and Whangbo[34], where they propose a fine-grained facial segmentation network based on a pyramid model. They demonstrated how multi-scale feature extraction methods can improve the efficacy of target detection.

Recently, the application of the Transformer model[35] has garnered increasing attention. Its core principle is to utilize self-attention mechanisms to capture interdependencies among different positions within the input sequence, thereby effectively facilitating the interaction and integration of information across various locations. Compared to traditional recurrent neural networks (RNNs), the Transformer allows for parallel processing of the entire input sequence, significantly enhancing both training and inference efficiency. The Transformer model consists of two primary modules: the encoder and the decoder. The encoder is responsible for extracting features from the input sequence, while the decoder is employed for generating outputs. Each encoder and decoder module is composed of multiple identical layers, typically six in number. A layer in the encoder comprises two main components: self-attention mechanism and feedforward neural network. The self-attention mechanism computes the relevance between each position in the sequence and all other positions, thus efficiently capturing global dependencies. The feedforward neural network then processes these features further, employing nonlinear mappings to elevate the expressive power of the model. The structure of the decoder mirrors that of the encoder but introduces an additional cross-attention mechanism between the self-attention mechanism and the feedforward network. This cross-attention module leverages the output of the encoder to guide the generation process of the decoder. By querying the correlations between the internally generated self-attention results within the decoder and the feature representations from the encoder, effective decoding of the features is achieved.

To mitigate these challenges, the Detection Transformer (DETR)[36] marked a paradigm shift in object detection by adopting the Transformer architecture. DETR integrates a CNN backbone with the Transformer, employs learnable queries to identify the image features encoded by the Transformer, executes set-based bounding box predictions through bipartite matching[37], obviates the need for manually defined components, implements a genuine form of non-maximum suppression, accelerates detection times, and exhibits robust performance on the MS COCO dataset[38]. Nonetheless, when directly applying the DETR model to the task of brand element detection, its accuracy and precision fall short of expectations. While general object detection methods exhibit remarkable performance in natural image detection, they often encounter limitations when dealing with brand images, such as: (1) the diverse and complex backgrounds of brand images can easily affect the detection precision of existing models; (2) the presence of variously shaped textual, iconographic, and graphic elements in brand images increases the complexity of the detection task; (3) small elements within brands tend to be of diminutive size, leading to suboptimal performance in detecting these tiny targets by current models; (4) the requirement for recognizing highly similar elements in brand detection tasks poses challenges for existing methods in capturing subtle differences.To address these issues, this paper proposes a multi-scale brand element detection network based on DETR, denoted as MSTED-Net. This model enhances the ability to detect small targets in brand detection by introducing a multi-scale feature enhancement pyramid (MFA-FPN) and a dual-perception attention module (DPAM). MSTED-Net is specifically tailored for brand detection tasks and is capable of effectively capturing details and multi-scale features present in brand images, thereby improving the model's detection precision and robustness. Rigorous testing and comparative analysis demonstrate that MSTED-Net outperforms existing solutions, providing a more accurate and precise solution for identifying brand elements.

Methodology

MSTED-Net, constructed upon the DETR architecture, consists of four primary components: Backbone, Neck, Transformer layers, and Prediction Heads, as illustrated in Fig. 2. The Dual Perception Attention Module (DPAM) is integrated into the Backbone network of MSTED-Net. DPAM extracts rich feature representations through its attention mechanism, emphasizing key features. This supplies critical high-resolution details

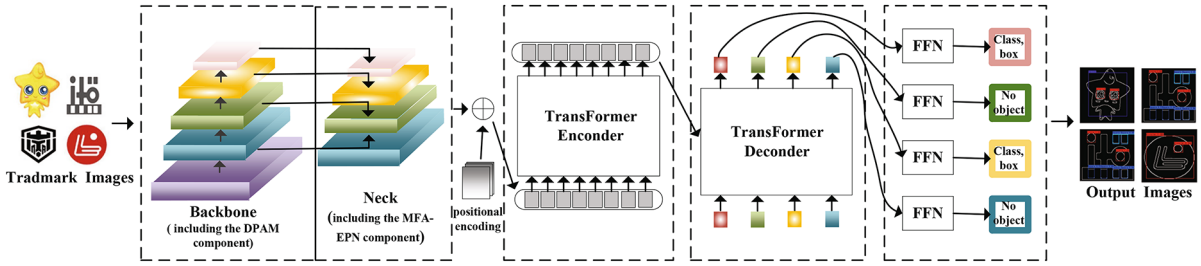


Figure 2. The architecture of MSTED-Net.

and broader contextual information, indispensable for detecting subtle and varied trademark elements, thus enriching the comprehension of complex scenes. Moreover, the Neck network integrates the Multi-scale Feature Augmentation Pyramid (MFA-FPN) to facilitate multi-scale feature fusion. Following processing through the MFA-FPN, the model produces a rich feature map encompassing multi-scale spatial hierarchies. Subsequently, this feature map, in conjunction with positional encoding, is inputted into the Transformer for encoding and decoding operations. This additional step refines feature representations, capturing profound dependencies and interactions among different trademark elements. Ultimately, the Prediction Heads deliver the predicted categories and location information of the targets.

Dual perception attention

In this model, the feature extraction layer utilizes ResNet50[39] to extract features from trademark images while reducing their dimensions. The detection performance of MSTED-Net depends significantly on the quality of the input features. To improve the feature extraction capabilities of the backbone network and efficiently detect element patterns within trademark images with high accuracy, a novel attention mechanism called the Dual Perception Attention Module (DPAM) is presented. This module enables subsequent layers and the final prediction heads to access more informative and relevant features by enhancing feature quality at an earlier stage. By addressing crucial concerns like the prevalence of small targets and complexities associated with image elements in trademark detection, DPAM effectively enhances the overall performance. DPAM is integrated into every stage of ResNet50, which improves the spatial and channel features within the feature maps, thus effectively enhancing target localization accuracy. A visual representation of the ResNet50 feature extraction network with embedded DPAM is provided in Fig. 3.

DPAM integrates SAM[40] and GCnet[41] to develop a dual-fusion attention mask that simultaneously captures both spatial and channel information. By utilizing global context aggregation in GCnet, DPAM can capture distant dependencies within the image, which is essential for comprehending the comprehensive structure and arrangement of trademark images. Conversely, SAM prioritizes the spatial aspects of the image, highlighting critical spatial information about features, thus noticeably amplifying local features within the frame. This distinctiveness enhances its efficacy in detecting and recognizing intricate and minor target components during pattern feature extraction tasks, offering substantial technical assistance for trademark element identification. The innovative fusion of these two attention mechanisms enables the model to dynamically adjust feature maps, generating more distinctive feature representations. Moreover, this integration allows the model to acquire a more abundant feature representation, transcending single informational constraints, ultimately leading to improved detection precision for complicated elements in trademarks. The architecture of DPAM is delineated in Fig. 4.

Different from typical approaches that solely involve stacking or serial connection of attention modules, DPAM performs a partitioning of the feature map, denoted as $M \in R^{C\times H\times W}$. This division is executed by dividing the feature map into two primary components: $M_c \in R^{C\times H\times W}$, which represents the input to the GCNet, and $M_s \in R^{C\times H\times W}$, which signifies the input to the SAM. The process of calculating the split is detailed as Eq. 1.

$$M_{\mathrm{s}} = \frac{1}{C} \mp \sum_{c=1}^{C} M_{chw} \tag{1}$$

Next, the features enhanced by both SAM and GCNet are combined across various dimensions. This fusion operation is carried out using Eq. 2.

$$M' = M_{\mathrm{s}1} \otimes M_{c1} \tag{2}$$

where M' represents the fused feature map, while $M_{\mathrm{s}1}$ and M_{c1} are the feature maps enhanced by SAM and GCNet, respectively.

Subsequently, the newly generated fused feature map is divided into two distinct scale feature maps based on our previous strategy. These scaled feature maps are then fused with the previously computed feature maps via

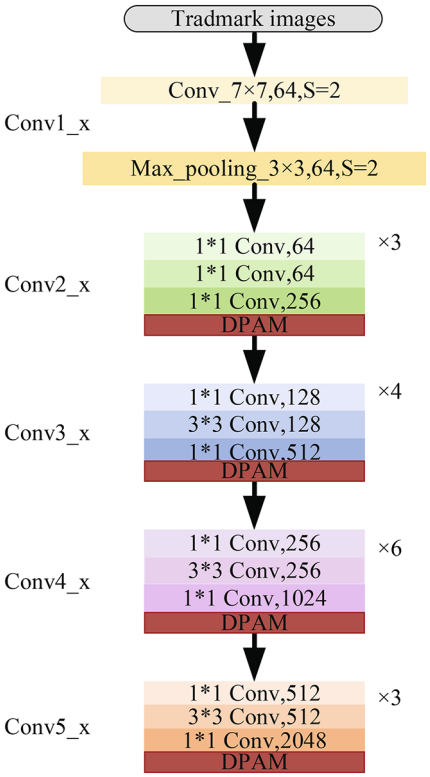


Figure 3. Feature extraction module.

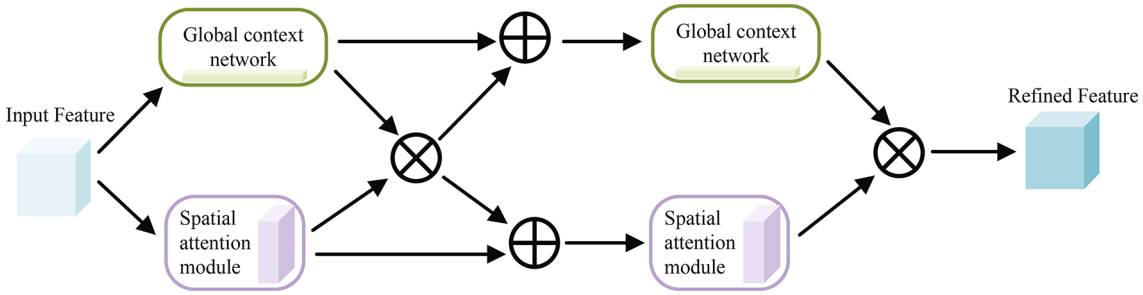


Figure 4. The structure of DPAM.

element-wise addition. After this fusion step, the resulting merged feature maps are passed onto two separate attention mechanisms for further processing. The computation of this fusion operation is described as:

$$M_{c2} = M'_2 \oplus M_{c1} \tag{3}$$

$$M_{s2} = M'_1 \oplus M_{s1} \tag{4}$$

where M_{c2} is the feature map input to the GCNet, and M_{s2} is the feature map input to the SAM. M'_1 and M'_2 are the two feature maps obtained by subdividing .

Ultimately, the two feature maps enhanced by the attention mechanisms are element-wise multiplication to obtain the final output feature map.

Multi-scale feature augmented pyramid

In trademark image processing, visual objects are frequently small and show substantial scale disparities, which may cause the feature extraction module to overlook or misinterpret crucial details while dealing with intricate and varying elements. To address this issue, utilizing a single feature map derived from the final feature extraction layer might result in the loss of fine-grained feature information of trademark components. Thus, we embrace the FPN[42] architecture for merging the four feature maps produced by the feature extraction layers and propose

MFA-FPN. MFA-FPN incorporates a Balancing Semantic Feature Pyramid Network (BS-FPN) and an Enhanced Feature Module (EFM). By seamlessly integrating diverse feature information, MFA-FPN effectively mitigates the discrepancy between information resolution and receptive fields across various layers, thereby improving the feature recognition capacity of trademark images. Moreover, through the combination of max-pooling and unpooling operations[43], MFA-FPN transmits high-level semantically rich features to lower levels, guaranteeing a consistent resolution for each feature map prior to fusion. This comprehensive integration of supplementary feature information eliminates information redundancy. The structure of MFA-FPN is depicted in Fig. 5.

Balanced semantic feature pyramid network

This research builds upon the original FPN by employing depth-integrated balanced semantic features to enhance multi-level features[44]; hence, the BS-FPN is introduced. The fundamental principle behind this approach lies in the integration of features from diverse depths into features with balanced semantics, which, in turn, improves the representation of multi-scale features. As demonstrated in Fig. 5, the three feature maps, F2, F3, and F5, derived from the feature extraction module are initially rescaled through up-sampling and down-sampling to ensure that their sizes align with the F4 feature map. Following this, an information aggregation operation is conducted. This process effectively mitigates the issue of feature information mismatch, resulting in a significant enhancement of the fusion effect. The detailed formula for the information aggregation operation is provided as Eq. 5.

$$C = \frac{1}{L} \sum_{l=l_{\min}}^{l_{\max}} C_l \tag{5}$$

where C represents the fused feature layer, C_l denotes the feature layer of the l-th level, and Lsignifies the total number of feature layers used. $l_{\min}$ stands for the lowest level feature, while $l_{\max}$ denotes the highest level feature.

Upon completion of the information aggregation operation, we implement a comparable yet opposite procedure to disseminate the accumulated features. This process augments the expressive capacity of the primal features, mitigate feature information disparity to some extent, and consequently strengthen the fusion outcome.

Enhanced feature module

In the conventional FPN architecture, 1×1 convolutions are frequently employed in the lateral connections. Nonetheless, this strategy presents several constraints. As the number of feature channels diminishes, spatial information tends to be lost, and the restricted receptive field of 1×1 convolutions impacts feature extraction. To overcome these challenges, we aim to expand the receptive field by incorporating pooling layers and elevating the convolutional kernel size. However, the pooling process sacrifices resolution, resulting in the loss of feature detail, whereas upping the convolutional kernel parameters might lead to model overfitting. Addressing this issue, we introduce the Enhanced Feature Module (EFM), whose structure is depicted in Fig. 6. This module fuses 1×1 and 3×3 convolution operations in the lateral connections, enabling the model to capture comprehensive feature information by leveraging the fusion of varying receptive fields. The combined features also simplify the

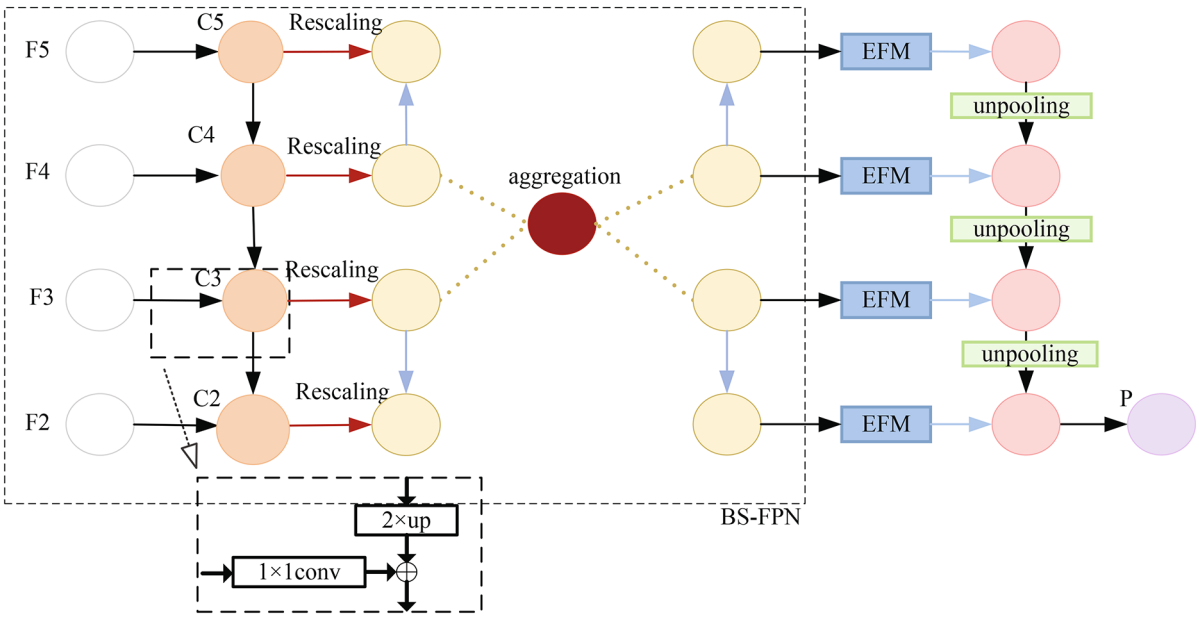


Figure 5. The structure of MFA-FPN.

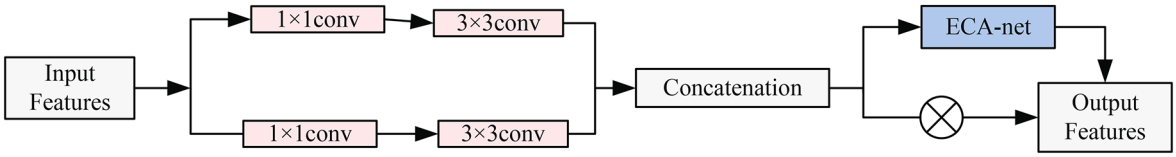


Figure 6. The structure of the EFM.

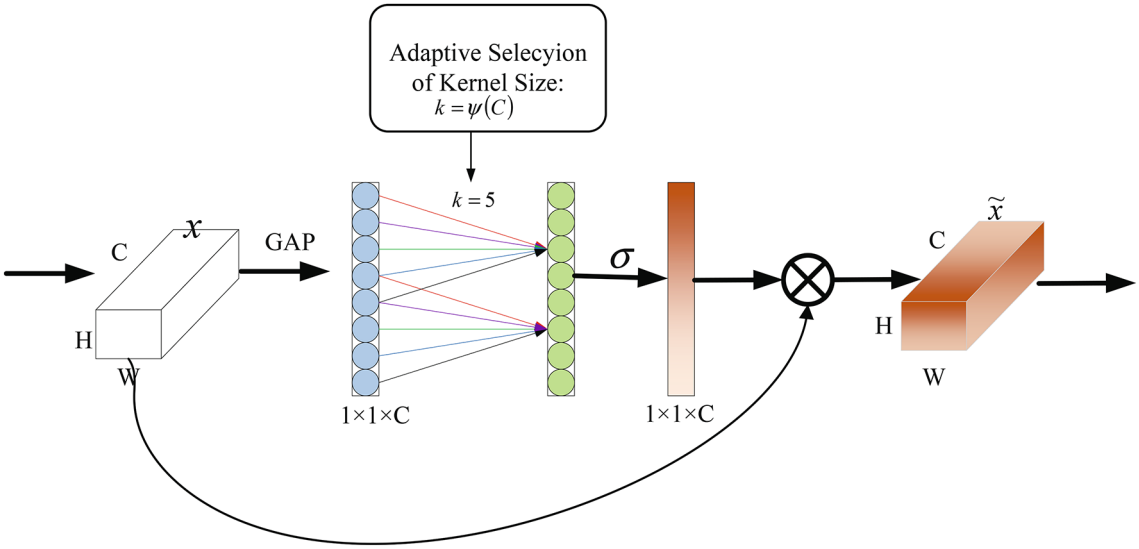


Figure 7. The structure of ECA-Net.

task of detecting small targets by the model. Moreover, maintaining a uniform number of feature channels across all layers ensures the efficacy and smoothness of subsequent feature fusion. This module significantly amplifies both feature extraction and integration, providing a robust foundation for enhanced model performance.

Alternatively, we introduced the Efficient Channel Attention Network (ECA-Net)[45] to facilely modify features adaptively. As delineated in Fig. 7, the initial input feature map undergoes global average pooling to foster global contextual interplay. Subsequently, the dimensions of the adaptive convolution kernel are determined using the subsequent formula:

$$C = \varphi(k) = 2^{(\gamma \times k - b)} \tag{6}$$

$$k = \psi(C) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{7}$$

where C represents the number of input channels; b and γ are set to 1 and 2, respectively; and $|t|_{odd}$ denotes the nearest odd number to t.

Subsequently, a single-dimensional convolution (Conv1D) is employed to determine the weights of the channels, with the Sigmoid activation function employed for weight normalization. Ultimately, the acquired weight values are assigned to the original input features via multiplication, enabling the network to concentrate on distinctive trademark element features while suppressing background noise.

After processing through ECM, the four feature maps experience a reduction in resolution. To facilitate the transfer of high-level semantic information-rich features to lower layers and ensure consistent resolutions among all feature maps prior to fusion, upsampling operations are employed. This approach allows for the full integration of complementary information-laden features, thereby avoiding redundancy. The current FPN (Feature Pyramid Network) employs nearest neighbor interpolation, which, although capable of retaining semantic information to a great extent, tends to lose precise positional relationships between feature maps after convolutional processing, making it challenging to exchange information effectively through simple summation. Therefore, this paper combines max-pooling and UnPooling upsampling operations[46], which largely reduces the loss of weight information and further conserves memory space. The UnPooling structure is illustrated in Fig. 8. First, max-pooling is used to select features at lower levels, preserving the position information of maximum values. Subsequently, these position informations are combined to perform the UnPooling operation, expanding the feature map while setting the remaining values to zero.

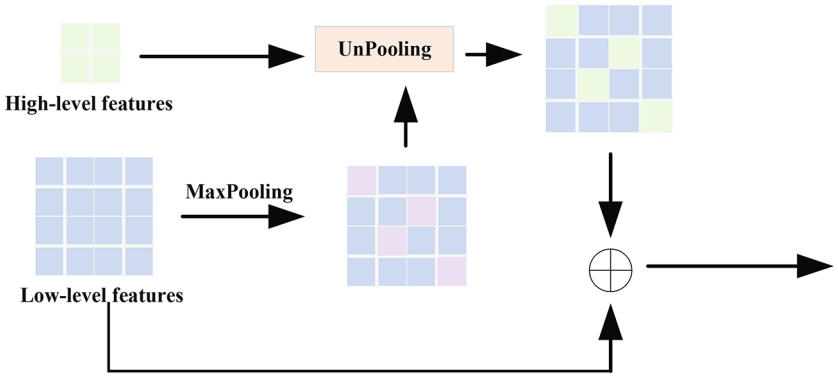


Figure 8. The flowchart of Unpooling operation.

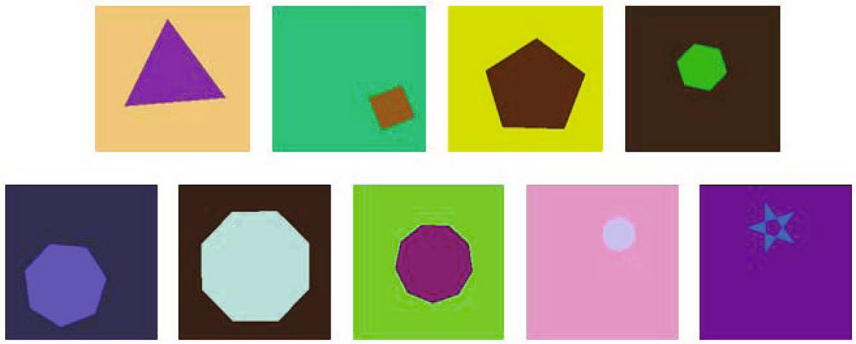


Figure 9. Examples from the 2D geometric shapes dataset.

Experiments

In this paper, we employ the open-source 2D Geometric Shapes Dataset[47] for the purpose of experimental investigation. This dataset comprises 90,000 labeled images, typically employed for geometric shape detection and classification tasks. It includes nine geometric shapes, each randomly generated onto 200x200 RGB images with random background colors and fillings. Additionally, the graphical positions and rotation angles are also randomly assigned. There are 10,000 images for each shape, totaling 90,000 labeled images. The dataset is divided into nine categories corresponding to unique geometric shapes (triangle, rectangle, pentagon, hexagon, heptagon, octagon, nonagon, circle, and star). The dataset is split into training and testing subsets in a ratio of 70% to 30%, respectively, for use with MSTED-Net, ensuring sufficient model training and generalization evaluation for the geometric shape detection task. An instance from the 2D Geometric Dataset is displayed in Fig. 9.

The complete experimental setup and configurations are summarized as follows: The host system is a Ubuntu 18.4 server, equipped with an NVIDIA GeForce RTX 3090 GPU featuring 24GB of VRAM. Python version 3.8 is adopted, along with the DETR deep learning framework and PyTorch 1.9.1 installed. ResNet50 is chosen as the backbone network, and the pre-trained DETR weights are utilized for training. The batch size is set to 2, the learning rate to 1e-4, and the AdamW optimizer is implemented. The training process lasts for 200 epochs.

The assessment metrics employed in this research are the Mean Average Precision (mAP), Recall, and Precision, which are commonly utilized in object detection tasks. Precision denotes the proportion of accurately predicted bounding boxes; Recall refers to the proportion of actual bounding boxes that are correctly predicted; and mAP represents the average precision across all categories. The respective formulas are defined as:

$$\mathrm{Pr}\,ecision = \frac{TP}{TP + FP} \times 100\% \tag{8}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \times 100\% \tag{9}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{10}$$

where *TP* represents true positives, *FP* shows false positives, *FN* indicates false negatives, *N* denotes the number of categories, and *i* indicates the value for the corresponding category.

Shapes	Faster R-CNN	SSD	YOLOv8	RT-DETR	DETR	MSTED-Net
Circle	86.15	100	100	100	100	100
Heptagon	88.94	100	100	99.2	99.7	100
Hexagon	94.26	100	100	100	100	100
Nonagon	87.5	99.06	100	100	100	100
Octagon	88.89	98.98	99.8	97.2	100	100
Pentagon	91.44	97.66	100	100	100	100
Square	83.03	100	100	100	99.7	99.9
Star	77.53	100	100	100	100	100
Triangle	82.57	100	100	100	99.9	100
Average	86.7	99.52	99.97	99.6	99.92	99.99

Table 1. Comparison of recall values (%) for different networks on the 2D geometric shapes dataset.

Shapes	Faster R-CNN	SSD	YOLOv8	RT-DETR	DETR	MSTED-Net
Circle	55.26	88.96	62.4	61.6	100	100
Heptagon	98.4	88.97	65.6	64.7	97.5	100
Hexagon	89.95	90.02	65.5	64.8	99.4	99.5
Nonagon	81.95	87.33	59.6	58.6	98.6	98.9
Octagon	94.85	88.98	61.5	61.2	100	99.8
Pentagon	93.44	90.76	65.2	65.1	99.5	99.1
Square	80.8	91.22	67.5	67.1	99.4	99.5
Star	73.16	77.85	70.6	70.5	99.3	99.4
Triangle	90.45	88.07	63.1	62.6	98.2	98.8
Average	84.25	88.02	64.6	64	99.1	99.44

Table 2. Comparison of precision values (%) for different networks on the 2D geometric shapes dataset.

Comparison experiments

To thoroughly evaluate the efficacy and dependability of the proposed method in trademark element detection, six detection approaches, Faster R-CNN, SSD, YOLOv8, RT-DETR, DETR, and MSTED-Net- were benchmarked on the 2D Geometric Shapes Dataset, and guaranteeing a consistent experimental environment and dataset. The recall values of the contrasting networks are reported in Table 1. In Table 1, MSTED-Net registered recall rates of 100%, 100%, 100%, 100%, 100%, 100%, 99.9%, 100%, and 100% across various geometric shape categories. The overall recall rate stands at 99.99%. In comparison to Faster R-CNN, SSD, YOLOv8, RT-DETR, and DETR, MSTED-Net boosts the overall recall by 13.29%, 0.47%, 0.02%, 0.39%, and 0.07%, respectively.

The precision values of the various networks are presented in Table 2. As observed in Table 2, MSTED-Net recorded precision rates of 100%, 100%, 99.5%, 98.9%, 99.8%, 99.1%, 99.5%, 99.4%, and 98.8% across different geometric shape categories. The overall precision rate amounts to 99.4%. In comparison to Faster R-CNN, SSD, YOLOv8, RT-DETR, and DETR, MSTED-Net enhances the overall precision by 15.59%, 11.42%, 34.84%, 35.44%, and 0.34%, respectively.

The ARP values of different networks are presented in Table 3. As depicted in Table 3, MSTED-Net achieved detection accabilities of 91.215%, 93.34%, 88.65%, 91.01%, 89.88%, 92.78%, 89.79%, 91.87%, and 91.56% across different geometric shape categories, outperforming other techniques in terms of detection accuracy for these shapes. Furthermore, the mean Average Precision (mAP) is 91.12%, which shows improvement of 5.11%, 4.92%, 21.72%, 24.52%, and 2.54% compared to Faster R-CNN, SSD, YOLOv8, RT-DETR, and DETR, respectively. These results indicate that MSTED-Net effectively tackles the challenges in trademark element detection. First, by incorporating DPAM, the network's capability to model deformed targets is strengthened, and its feature representation capability is optimized. Second, the construction of the MFA-FPN effectively integrates multi-scale features, addressing scale variation issues more efficiently and further enhancing detection accuracy.

On the 2D geometric shapes dataset, the mean Average Precision (mAP) values of six algorithms including Faster R-CNN, Single Shot MultiBox Detector (SSD), YOLOv8, Region-based Target Detection with Reinitialization (RT-DETR), Detection Transformers (DETR), and Minimum Spanning Tree Graph Embedding (MSTED-Net) are compared after 200 training epochs, as illustrated in Fig. 10. The graph clearly demonstrates that the proposed algorithm outperforms other methods in terms of accuracy.

From the data in Table 4, it is evident that although the MSTED-Net achieves an inference speed of 52.6 milliseconds, which may seem slower compared to other models, its application value in real-time brand monitoring cannot be overlooked. In practical brand surveillance scenarios, detection accuracy is often more critical than speed because precise identification of brand elements can effectively prevent infringement activities

Shapes	Faster R-CNN	SSD	YOLOv8	RT-DETR	DETR	MSTED-Net
Circle	84.0	87.9	65.2	67.9	88.3	91.2
Heptagon	92.0	86.3	72.3	66.1	88.3	93.3
Hexagon	92.0	87.6	69.8	67.1	88.4	88.7
Nonagon	90.0	88.3	66.1	60.1	88.6	91.0
Octagon	89.0	87.9	63.2	66.3	88.7	89.9
Pentagon	87.0	82.3	73.5	70.0	88.3	92.8
Square	85.0	86.6	70.3	70.6	88.9	89.8
Star	76.0	79.8	76.6	71.4	88.9	91.9
Triangle	79.0	89.3	67.7	69.5	88.88	91.6

Table 3. Comparison of AP values (%) for different networks on the 2D geometric shapes dataset.

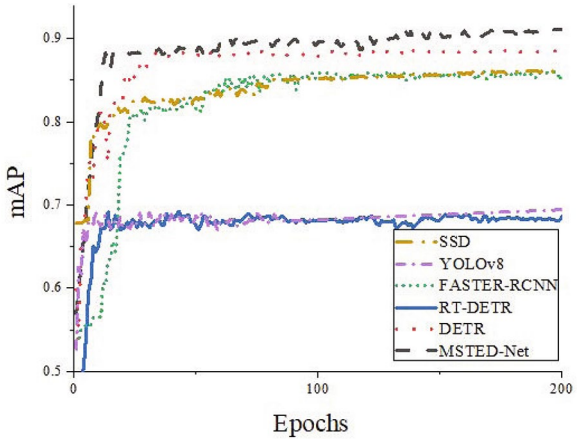


Figure 10. The mAP curves.

Methods	Faster R-CNN	SSD	YOLOv8	RT-DETR	DETR	MSTED-Net
Speed (ms)	47.4	33.8	8.6	16.7	35.2	52.6

Table 4. A comparison of inference speed among different models.

and protect corporate intellectual property rights. In real-time monitoring environments, despite being less rapid than models like YOLOv8, the higher accuracy of MSTED-Net ensures effective recognition of trademarks in complex settings. This is crucial for legal compliance and brand protection. Moreover, plans are underway to further optimize the model to enhance its inference speed without compromising accuracy. Such improvements will help bolster MSTED-Net's competitiveness in real-time applications, making it a stronger candidate for deployment where both speed and precision are important.

As depicted in Table 4, the first image consists of a single star and two circular shapes. Neither Faster R-CNN nor YOLOv8 were able to successfully detect all three targets, leading to substantial false positives. In the second image, there are two circular shapes and six rectangular shapes. Faster R-CNN identified two circles and three rectangles, however, it overlooked the remaining three rectangles, indicating missed detections. Despite detecting all the targets with YOLOv8, this model also mistakenly classified other non-target objects. Lastly, in the third image containing two hexagonal shapes and two square shapes, Faster R-CNN failed to identify any targets properly, whereas YOLOv8 could only detect one target, showcasing both false and missed detections.

On the other hand, MSTED-Net, featuring an innovative attention mechanism and pyramid module, prioritizes crucial features within the image, allowing for clear and precise identification of all geometric elements with reduced false and missed detection rates. Moreover, MSTED-Net surpassed the other two models in accurately detecting smaller targets in trademark images, illustrating its remarkable advantages.

Ablation experiments

To evaluate the efficiency of the DPAM, BS-FPN, EFM, and unpooling operation in MSTED-Net for trademark element detection tasks, we conducted five ablation experiments:

- Employing the DETR network with Resnet50 serving as the main trunk feature extraction network;

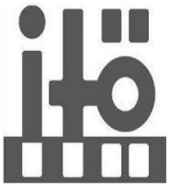
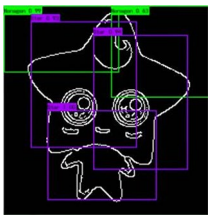
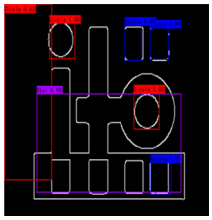
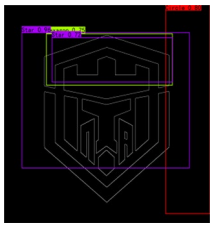
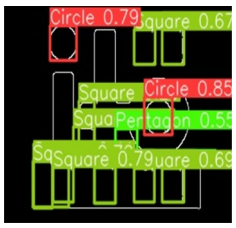
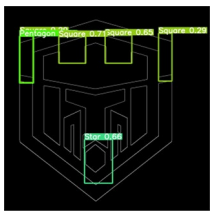
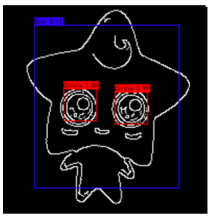
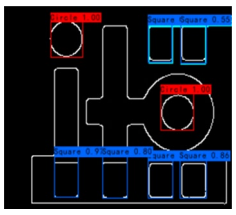
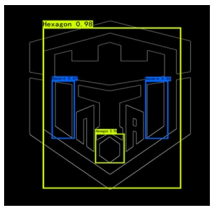
Methods	Example 1	Example 2	Example 3
Original Images			
Faster R-CNN			
YOlOv8			
MSTED-Net			

Table 5. Comparison of trademark element detection results.

Cases	DPAM	BS-FPN	EFM	UnPooling	mAP/%	Recall/%	Precision/%
1					88.58	99.92	99.1
2	✓				90.32	99.95	99.23
3	✓	✓			91.07	99.97	99.38
4	✓	✓	✓		91.10	99.98	99.40
5		✓	✓	✓	91.11	99.98	99.42
6	✓	✓	✓	✓	91.12	99.99	99.44

Table 6. Results of ablation experiments.

- Incorporating the dual perception attention module into the trunk feature extraction network of Case 1;
- Adding a balanced semantic feature pyramid to the neck network following the Resnet50 in Case 1;
- Integrating the enhanced feature module based on Case 3;
- Introducing MFA-FPN into the DETR model;
- Appending an upper pooling feature fusion module to Case 4.The effectiveness of these modifications was assessed by examining the Recall, mAP (mean Average Precision), and Precision metrics. The detection outcomes from various improvement techniques applied to the test set are presented in Table 6.

By comparing Case 1 and Case 2, it becomes apparent that integrating the DPAM into the backbone boosts the mAP by 1.74%, Recall by 0.03%, and Precision by 0.13%. This indicates that the DPAM enhancements help the model to concentrate on crucial features, effectively avoiding the loss of smaller target features and enhancing its capability to extract them. Comparing Case 1 and Case 3, the inclusion of the BS-FPN compiles features of varying magnitudes, ensuring comprehensive utilization of all scale features. Consequently, there is a 0.75% rise in mAP, 0.02% in Recall, and 0.15% in Precision. By combining the EFM with the BS-FPN via convolutional operations that fuse receptive fields of diverse sizes and introducing the ECA-Net for adaptive feature adjustment, there is a slight 0.03% uptick in mAP, a 0.01% increment in Recall, and a 0.02% rise in

Precision. Lastly, applying the UpPooling operation to the pyramid framework yields a 0.02% increase in mAP, a 0.01% rise in Recall, and a 0.04% increase in Precision. When compared to Cases 1, 2, 3, and 4, it is evident that merging all four components significantly surpasses the individual additions of each component. The mAP escalates by 2.54%, Recall improves by 0.07%, and Precision increases by 0.34% relative to the baseline model, showcasing a considerable advancement in the model's detection capabilities.

Discussion

The proposed MSTED-Net demonstrates exceptional performance in brand element detection, thanks to the innovative combination of the Dual Perceptual Attention Module (DPAM) and the Multi-Scale Feature Enhancement Pyramid (MFA-FPN), which enhances the ability of the DETR model to fuse multi-scale features. Consequently, the precision has significantly improved, showcasing the advantages of this method in brand detection. Nonetheless, computational cost and inference speed remain challenges, especially when deploying on resource-constrained devices. The trade-off between increasing accuracy and reducing inference speed may limit the applicability of MSTED-Net in real-time systems.

In future work, to address the issue of heavy computation and slow speed, we plan to employ pruning and quantification technologies to reduce the complexity and computational demands of the model. Alternatively, we might consider introducing lightweight convolutional networks on top of the existing model to increase inference speed. We will also explore the integration of reinforcement learning to apply dynamic adaptation strategies to assist the model in adapting to new detection strategies. This integration could potentially improve the model's generalizability to new datasets. Further enhancements to the Transformer layers are still possible, particularly in extracting and refining granular features. This not only improves detection accuracy but also aids the model in better generalizing across varying object sizes. Additionally, we will investigate domain adaptation techniques to make MSTED-Net more adaptable to different national or regional brand datasets. Trademark laws and symbols vary across legal systems, and MSTED-Net's adaptation via domain adaptation technology will significantly enhance its global applicability.

The MSTED-Net architecture can also be extended to other small object detection tasks. Potential applications include satellite imagery (for detecting small objects such as vehicles or buildings), medical imaging (for detecting anomalies such as tumors or lesions), and fine-grained image recognition (for distinguishing highly similar objects). In these domains, the ability to perform high-precision small object detection is crucial, making MSTED-Net a valuable tool beyond its original purpose.

Conclusion

This paper presents MSTED-Net, a multi-scale trademark element recognition network, aiming to address graphic elements within trademarks.DPAM is integrated into the backbone network, enabling the model to integrate various attention mechanisms and capture rich spatial and contextual information, thereby alleviating the problem of feature loss resulting from reliance on a single mechanism. Moreover, the MFA-FPN is developed to strengthen interactions among different scale features, augment the semantic depth of shallower features, and effectively enhance feature representation abilities. Experimental findings indicate that MSTED-Net outperforms the original DETR algorithm in terms of mAP by 2.54%, while simultaneously elevating accuracy and recall rates by 0.34% and 0.07%, respectively. These accomplishments not only validate the algorithm's efficacy in reducing element detection errors but also underscore its potential to boost detection accuracy, exhibiting strong performance in trademark element recognition tasks. In our future research, we plan to develop more sophisticated element detection networks and continually refine the training procedure to tackle the difficulties inherent in trademark element detection.

Data availability

Data sets generated during the current study are available from the corresponding author on reasonable request.

Received: 13 July 2024; Accepted: 4 November 2024
Published online: 25 November 2024

References

1. Dixit, U., Shirdhonkar, M. & Sinha, P. G. Automatic logo detection from document image using HOG features. *Multimedia Tools Appl.* **82**, 1–16. https://doi.org/10.1007/s11042-022-13300-5 (2022).
2. Wu, X., Sahoo, D. & Hoi, S. Recent advances in deep learning for object detection. *Neurocomputing* **396**, 39–64. https://doi.org/10.1016/j.neucom.2020.01.085 (2020).
3. Fehérvári, I. & Appalaraju, S. Scalable logo recognition using proxies. In *2019 IEEE Winter Conference on Applications of Computer Vision*, 715–725. https://doi.org/10.1109/WACV.2019.00081 (IEEE, Hawaii, 2019).
4. Truong Hoang, V. Vehicle logo recognition using HOG descriptor and sparsity score. *Telecommun. Comput. Electron. Control* **18**, 3019–3025. https://doi.org/10.12928/telkomnika.v18i6.16133 (2020).
5. Bianco, S., Buzzelli, M., Mazzini, D. & Schettini, R. Logo recognition using cnn features. In *18th International Conference on Image Analysis and Processing*, vol. 9280, 438–448, https://doi.org/10.1007/978-3-319-23234-8_41 (Springer, Genoa, Italy, 2015).
6. Zou, Z., Chen, K., Shi, Z., Guo, Y. & Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **111**, 257–276. https://doi.org/10.1109/JPROC.2023.3238524 (2023).
7. Sahel, S., Alsahafi, M., Alghamdi, M. & Alsubait, T. Logo detection using deep learning with pretrained CNN models. *Eng. Technol. Appl. Sci. Res.* **11**, 6724–6729. https://doi.org/10.48084/etasr.3919 (2021).
8. Hou, S. et al. Deep learning for logo detection: A survey. *ACM Trans. Multimed. Comput. Commun. Appl.* **20**, 1–23. https://doi.org/10.1145/3611309 (2023).
9. Montserrat, D., Lin, Q., Allebach, J. & Delp, E. Logo detection and recognition with synthetic images. *Electron. Imaging* **3371–3377**, 2018. https://doi.org/10.2352/ISSN.2470-1173.2018.10.IMAWM-337 (2018).

10. Alsheikhy, A., Said, Y. & Barr, M. Logo recognition with the use of deep convolutional neural networks. *Eng. Technol. Appl. Sci. Res.* **10**, 6191–6194. https://doi.org/10.48084/etasr.3734 (2020).
11. Li, X., Wei, T., Chen, Y., Tai, Y.-W. & Tang, C.-K. Fss-1000: A 1000-class dataset for few-shot segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2866–2875. https://doi.org/10.1109/CVPR42600.2020.00294 (IEEE, WA, USA, 2020).
12. Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 580–587. https://doi.org/10.1109/CVPR.2014.81 (IEEE, OH, USA, 2013).
13. Li, Z., Wang, F. & Wang, N. Lidar R-CNN: An efficient and universal 3D object detector. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7542–7551. https://doi.org/10.1109/CVPR46437.2021.00746 (IEEE, TN, USA, 2021).
14. Xie, X., Cheng, G., Wang, J., Yao, X. & Han, J. Oriented R-CNN for object detection. In *2021 IEEE/CVF International Conference on Computer Vision*, 3500–3509. https://doi.org/10.1109/ICCV48922.2021.00350 (IEEE, TN, USA, 2021).
15. He, K., Zhang, X., Ren, S. & Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824 (2014).
16. Wang, X., Wang, S., Cao, J. & Wang, Y. Data-driven based tiny-yolov3 method for front vehicle detection inducing SPP-Net. *IEEE Access* **8**, 110227–110236. https://doi.org/10.1109/ACCESS.2020.3001279 (2020).
17. Jeon, J., Jeong, B., Baek, S. & Jeong, Y.-S. Hybrid malware detection based on Bi-LSTM and SPP-Net for smart IoT. *IEEE Trans. Ind. Inf.* **18**, 4830–4837. https://doi.org/10.1109/TII.2021.3119778 (2022).
18. Girshick, R. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision*, 1440–1448, https://doi.org/10.1109/ICCV.2015.169 (IEEE, Santiago, Chile, 2015).
19. Rani, S., Ghai, D. & Kumar, S. Object detection and recognition using contour based edge detection and fast R-CNN. *Multimedia Tools Appl.* **81**, 42183–42207. https://doi.org/10.1007/s11042-021-11446-2 (2022).
20. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031 (2015).
21. Zeng, L., Sun, B. & Zhu, D. Underwater target detection based on faster R-CNN and adversarial occlusion network. *Eng. Appl. Artif. Intell.* **100**, 104190. https://doi.org/10.1016/j.engappai.2021.104190 (2021).
22. Avola, D. et al. Ms-faster r-cnn: Multi-stream backbone for improved faster r-cnn object detection and aerial tracking from uav images. *Remote Sens.* **13**, 1670. https://doi.org/10.3390/rs13091670 (2021).
23. Liu, W. et al. Ssd: Single shot multibox detector. In Leibe, B., Matas, J., Sebe, N. & Welling, M. (eds.) *2016 European Conference on Computer Vision*, 21–37 (Springer International Publishing, Cham, 2016).
24. Zhang, X., Zhang, Y., Gao, T., Fang, Y. & Chen, T. A novel SSD-based detection algorithm suitable for small object. *IEICE Transactions on Information and Systems* **E106.D**, 625–634. https://doi.org/10.1587/transinf.2022DLP0037.
25. Zhai, S., Shang, D., Wang, S. & Dong, S. DF-SSD: An improved SSD object detection algorithm based on densenet and feature fusion. *IEEE Access* **8**, 24344–24357. https://doi.org/10.1109/ACCESS.2020.2971026 (2020).
26. Jiang, P., Ergu, D., Liu, F., Cai, Y. & Ma, B. A review of YOLO algorithm developments. *Proc. Comput. Sci.* **199**, 1066–1073. https://doi.org/10.1016/j.procs.2022.01.135 (2022).
27. Mao, K., Jin, R., Chen, K., Mao, J. & Dai, G. Trinity-YOLO: High-precision logo detection in the real world. *IET Image Proc.* **17**, 2272–2283. https://doi.org/10.1049/ipr2.12791 (2023).
28. Diwan, T., Ani, A. & Tembhurne, J. Object detection using YOLO: Challenges, architectural successors, datasets and applications. *Multimed. Tools Appl.* **82**, 9243–9275. https://doi.org/10.1007/s11042-022-13644-y (2022).
29. Chen, Q. et al. You only look one-level feature. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13034–13043. https://doi.org/10.1109/CVPR46437.2021.01284 (IEEE, TN, USA, 2021).
30. Wang, C.-Y., Bochkovskiy, A. & Liao, H.-Y. M. YOVOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7464–7475, https://doi.org/10.1109/CVPR52729.2023.00721 (IEEE, BC, Canada, 2023).
31. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 318–327. https://doi.org/10.1109/TPAMI.2018.2858826 (2020).
32. Liu, S., Cai, T., Tang, X., Zhang, Y. & Wang, C. Visual recognition of traffic signs in natural scenes based on improved RetinaNet. *Entropy* **24**, 112. https://doi.org/10.3390/e24010112 (2022).
33. Cheng, X. & Yu, J. Retinanet with difference channel attention and adaptively spatial feature fusion for steel surface defect detection. *IEEE Trans. Instrum. Meas.* **70**, 1–11. https://doi.org/10.1109/TIM.2020.3040485 (2021).
34. Umirzakova, S. & Whangbo, T. K. Detailed feature extraction network-based fine-grained face segmentation. *Knowl.-Based Syst.* **250**, 109036. https://doi.org/10.1016/j.knosys.2022.109036 (2022).
35. Khan, S. et al. Transformers in vision: A survey. *ACM Comput. Surv.* **54**, 1–41. https://doi.org/10.1145/3505244 (2022).
36. Carion, N. et al. End-to-end object detection with transformers. In *2020 European Conference on Computer Vision*, 213–229. https://doi.org/10.1007/978-3-030-58452-8_13 (Springer, Glasgow, UK, 2020).
37. Meng, D. et al. Conditional DETR for fast training convergence. In *2021 IEEE/CVF International Conference on Computer Vision*, 3631–3640. https://doi.org/10.1109/ICCV48922.2021.00363 (IEEE, QC, Canada, 2021).
38. Lin, T.-Y. et al. Microsoft coco: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T. (eds.) *2014 European Conference on Computer Vision*, 740–755 (Springer International Publishing, Cham, 2014).
39. Loey, M., Manogaran, G., Taha, M. H. N. & Khalifa, N. E. M. Fighting against COVID-19: A novel deep learning model based on YOLOv2 with ResNet-50 for medical face mask detection. *Sustain. Cities Soc.* **65**, 102600. https://doi.org/10.1016/j.scs.2020.102600 (2021).
40. Zhu, X., Cheng, D., Zhang, Z., Lin, S. & Dai, J. An empirical study of spatial attention mechanisms in deep networks. In *2019 IEEE International Conference on Computer Vision*, 6687–6696, https://doi.org/10.1109/ICCV.2019.00679 (IEEE, Seoul, Korea, 2019).
41. Cao, Y., Xu, J., Lin, S., Wei, F. & Hu, H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *2019 IEEE International Conference on Computer Vision*, 1971–1980, https://doi.org/10.1109/ICCVW.2019.00246 (IEEE, Seoul, 2019).
42. Wu, Y. et al. Rethinking classification and localization for object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10183–10192, https://doi.org/10.1109/CVPR42600.2020.01020 (IEEE, WA, USA, 2020).
43. Lee, Y. & Park, J. Centermask: Real-time anchor-free instance segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13903–13912, https://doi.org/10.1109/CVPR42600.2020.01392 (IEEE, WA, USA, 2020).
44. Zhang, T. et al. Balanced feature pyramid network for ship detection in synthetic aperture radar images. In *2020 IEEE Radar Conference on Radar*, 1–5, https://doi.org/10.1109/RadarConf2043947.2020.9266519 (IEEE, Florence, Italy, 2020).
45. Wang, Q. et al. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11531–11539, https://doi.org/10.1109/CVPR42600.2020.01155 (IEEE, WA, USA, 2020).
46. Umirzakova, S. & Whangbo, T. K. Detailed feature extraction network-based fine-grained face segmentation. *Knowl.-Based Syst.* **250**, 109036. https://doi.org/10.1016/j.knosys.2022.109036 (2022).
47. Korchi, A. E. & Ghanou, Y. 2D geometric shapes dataset - for machine learning and pattern recognition. *Data Brief* **32**, 106090. https://doi.org/10.1016/j.dib.2020.106090 (2020).

Acknowledgements
This work was supported by the National Key Research and Development Program (No.2021YFC3340402).

Author contributions
LL: conception and design of work, analysis, statistical processing, visualisation and interpretation of data, drafting original manuscript, editing; XW:conception and design of work, analysis, data interpretation, reviewing and editing, acquisition of funds; WQY: analysis, review and editing, providing study materials and context information; All authors are accountable for own contributions and have reviewed and approve the submitted manuscript.

Declarations

Competing interests
Te authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to X.W. or W.Q.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.