[1]Tao Li

A Study of Multimodal Perception Mechanisms for Target Detection Accuracy Enhancement in Computer Vision

Journal of Electrical Systems

Abstract: - In this paper, a multimodal perception mechanism framework is constructed to comprehensively perceive complex environments by integrating different modal data. The target detection framework consists of five modules such as feature fusion block and feature enhancement, which fuse visual and textual features. Multi-scale text features are extracted by constructing line-level text embedding maps and converting them to 2D feature maps. During the feature extraction process, text features are incorporated multiple times to achieve deep fusion of multimodal features. In addition, attention mechanism and contextual information are introduced to optimize the target features and enhance the detection ability of complex scenes. The results show that the average accuracy of the multimodal perception mechanism is above 95%, and the equilibrium point is around 0.96. Under the color interference condition, the average accuracy of multimodal perception is 93.28, and the mAP is significantly improved to 95.4% when the attention mechanism and contextual information are introduced. A series of results verified that in computer vision target detection, the multimodal perception mechanism has the highest detection accuracy and achieved the best enhancement performance in the same period.

Keywords: multimodal perception mechanism; target detection; text features; attention mechanism; computer vision

1. Introduction

With the rapid development of artificial intelligence technology, research in the field of computer vision has made remarkable progress, and target detection, as one of the core tasks, has received widespread attention [1]. However, traditional target detection methods are mainly based on single-modal data, such as images or videos, which faces many challenges when dealing with complex scenes and diverse targets [2]. Therefore, the application of multimodal perception mechanisms in improving the accuracy of target detection is particularly important [3]. The core idea of multimodal perception mechanism is to realize a more comprehensive and accurate perception and understanding of the environment by fusing data from different perceptual modalities, such as visual, auditory, and tactile [4]. In the field of computer vision, this mechanism is particularly important. The multimodal perception mechanism is able to combine information from multiple modalities, such as images, text, sound, etc., in order to provide richer and more diverse input data [5]. This fusion of multimodal data not only increases the information source for target detection, but also improves the robustness and generalization ability of the model.

Li, Z and other scholars studied the 3D target detection technique in the artificial algorithm of computer vision and applied it to the research of intelligent driving cars [6]. Xu, C et al. to simplify the difficulty of 3D object detection, based on an improved 2D target detection method to achieve target detection on RGB images [7].

[1] [1]* School of Electronic And Electrical Engineering，Ningxia University Yinchuan 750021,Ningxia, China.Email:albert2178@163.com

Bengamra, S et al. showed that there is a large amount of information present in artwork images and in order to extract the required information in a digitized form, various computer vision detection methods have been investigated to improve the performance of artwork detection [8]. Cao, D et al. In order to maximize the recognition of categories and locations in a target image, the problem of poor performance of existing detection methods in detecting dense small objects is addressed. A convolutional neural network is applied to extract multi-scale features and combined with variable convolutional structure for geometric transformation. The study was validated and found to provide a significant improvement in accuracy in computer vision detection [9]. Tsai, T. H. et al. in their research on target detection in the computer field constructed a real-time monitoring system to simplify the complexity of the algorithm. By monitoring the location of the object in real time and labeling the target information to classify it, thus outperforming other detection methods in terms of computational efficiency and detection accuracy [10]. Duan, S et al. showed that the improvement of detection accuracy is a research priority in the field of computer vision and proposed the OC-Anchors detection technique. By creating object category tracing points and welcoming god category attributes, the average accuracy range was improved to 27.1% on the COCO dataset [11]. Muralidhara, S and other scholars established a video detection framework in computer vision target detection by applying a backbone network to extract key features and utilizing aggregated support frame features to improve target detection advances, verifying that average accuracies of 49.8 and 52.5 can be achieved [12]. Yang, W. J et al. proposed an improved motion target detection system in computer vision that combines an object detector and a long and short-term memory to improve motion target detection performance. For the vehicle information associated spatial features, it was concluded after a step-by-step evaluation that the proposed detection system was able to maintain the computational overhead and still accurately detect the target vehicle in a low level [13].

In conjunction with the above research, it was found that the task of target detection in the field of computer vision faces challenges when dealing with complex scenes and diverse targets. Traditional target detection methods are based on single modal data and have limitations. In this paper, a complete perceptual closed loop is formed by interconnecting the data flow and feedback mechanism, which can sense and understand the environment more comprehensively and accurately and provide richer information for target detection. Using the text feature extraction module and the feature fusion module, the complementary information of different modalities can be fully utilized to make the information in the feature map richer, so as to improve the accuracy and robustness of target detection. The deep fusion and complementarity of multimodal features is realized by introducing the attention mechanism and contextual information. The module uses the null convolution and attention mechanism to select multi-scale contextual features, which improves the diversity of features and the generalization performance of the model, improves the average accuracy of the target detection task, and still maintains high detection accuracy under different conditions.

2. Multimodal perception mechanism

The multimodal perception mechanism framework is a complex system that combines multiple sensors and data processing techniques to achieve multidimensional and comprehensive perception of the external environment [14]. The main components of the framework include multimodal data input, feature extraction, multimodal fusion, target detection and tracking, and long term memory and prior knowledge. These components are

interconnected through data flow and feedback mechanisms to form a complete perceptual closed loop.

Considering the feature space inconsistency of multimodal data, targeted feature extraction is performed on multimodal data and multichannel features are fused in the multimodal integration module [15]. The fusion process will combine the a priori knowledge from consecutive frames, which can enhance the cognitive ability of the whole system to the environment, and dramatically improve the accuracy and robustness of the algorithms of target detection, semantic segmentation, and target tracking in abnormal environments. The framework of the multimodal perception mechanism is shown in Fig. 1, which starts from multimodal data input, goes through the steps of feature extraction and multimodal fusion, and finally outputs the target detection and tracking results. In this process, data flows between different modules, forming a clear data flow. The sensing mechanism continuously optimizes the sensing results through the feedback mechanism. For example, when a target is detected, information such as the position and speed of the target can be fed back to the control system, which makes decisions based on this information and executes the corresponding operations through the actuators. At the same time, the control system can also feedback the execution results to the perception mechanism to help the perception mechanism better understand the external environment.

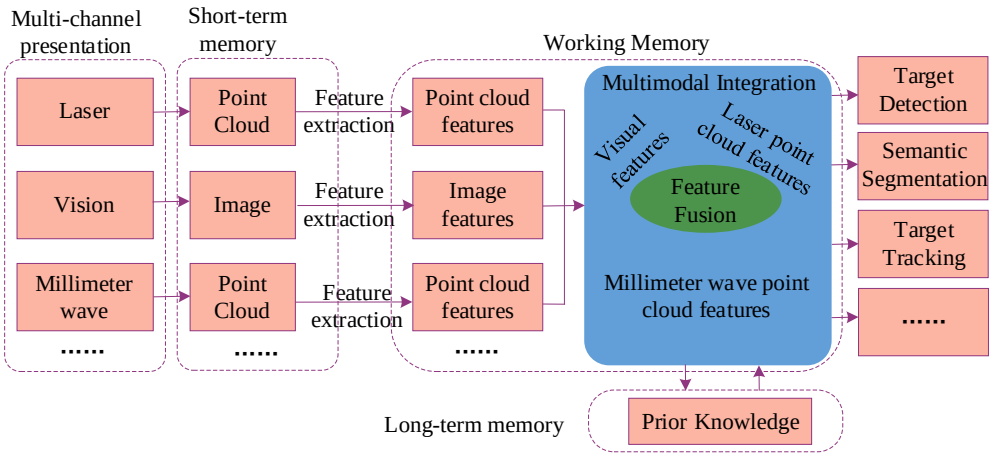


Figure 1 Multimodal perception mechanism framework

3. Target detection accuracy enhancement based on multimodal sensing mechanism

In computer vision target detection, multimodal perceptual networks enhance feature representation by fusing visual and textual features and utilizing feature enhancement modules, which require the processing of a large collection of data, and in this regard, this paper conducts a study for improving target detection accuracy.

3.1 Multi-target detection model architecture

3.1.1 Overall architecture

The architecture of the multimodal perception mechanism in computerized target detection is shown in Fig. 2, which consists of five modules: text feature extraction module, feature fusion module, feature enhancement module, feature pyramid network, and region generation network. Among them, the text feature extraction module mainly consists of four different convolutional and regularization layers, which are the basic components for performing text feature extraction operations. The feature fusion module is dominated by the ResNet network, which realizes the deep fusion of multimodal features through its powerful feature

representation capability and retains the rich feature information, thus making full use of the information of both. The feature enhancement module mainly consists of convolutional layers and up-sampling layers, similar in appearance to the feature pyramid network, which mainly realizes the transfer of the representation information of features of different scales over the channel, so that the low-level features also contain rich semantic information.The FPN transforms the feature maps of neighboring layers to the same size, and then sums up the corresponding positional elements, with the aim of transferring the strong semantic information in the high-level features to the low-level features, in order to It realizes the combination of low-level high-resolution information and high-level strong semantic information, so as to improve the detection performance.PRN is mainly composed of convolutional, intermediate, classification and regression layers, and its essence is to classify and regress the target region on the feature map based on the mechanism of sliding window and tracing frame, and to generate a series of candidate regions.

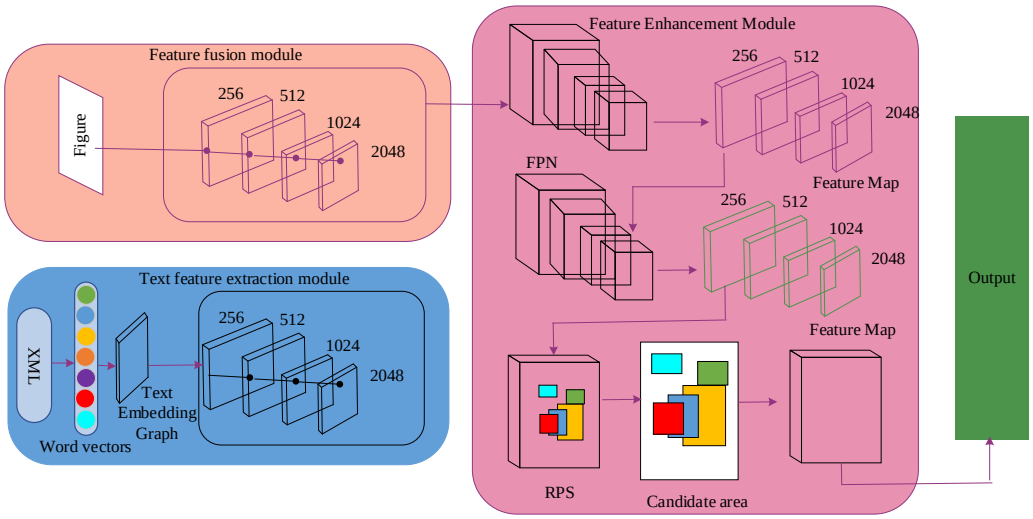


Figure 2 Multimodal perception mechanism target detection model

3.1.2 Text feature extraction module

Text features can distinguish visually similar regions and improve detection accuracy [16]. In order to utilize the text information, the text sequence needs to be converted from one-dimensional to two-dimensional, and in this paper, a line-level text embedding map is constructed [17]. Referring to the idea of sentence transformation, the text information in the document image is trained to get the corresponding word vectors, and the text information used to generate the word vectors is obtained by parsing the PDF file corresponding to the document image. Considering the overhead of the model and the reusability of the text information, the text information obtained from parsing is saved into XML files, and the word vectors obtained from training are also saved into the corresponding files. The text features are able to distinguish visually similar regions and contribute to the overall document obtained information contains line level text information and paragraph level text letter detection accuracy.

In order to utilize the text information, it is necessary to the text sequence interest, through a comparison algorithm to compare the two coordinate information, and thus build two columns from one-dimensional converted to two-dimensional, this paper builds the line level text embedding map. Considering the inclusion

relationship in the detection target, a hierarchical relationship data set is built in the training document image. If on image $x \in \mathbb{R}^{H \times W \times C_0}$, there is a set of rows of information text information to get the corresponding word vector, used to generate the word vector of text information is through I_l, where H and W are the height and width of the image, respectively, and C_0 represents the channel dimension of the image. The expression for the information of the rows is:

$$I_l = \left\{ \left(l_k, b_k \right) \middle| k = 1, \cdots, n \right\} \tag{1}$$

Where n is the total number of lines and the reusability of text information, the text information obtained from parsing is saved to the XML bibliography. l_k is the text information of line k, $b_k = \left(x_{ul}^k, y_{ul}^k, x_{br}^k, y_{br}^k \right)$ represents the coordinate box of the text in line k, and $\left(x_{ul}^k, y_{ul}^k \right)$ and $\left(x_{ul}^k, y_{ul}^k \right)$ represent the coordinates of the upper left and lower right corners, respectively. The formula for text embedding Figure $T \in \mathbb{R}^{H \times W \times C_0}$ is defined as follows:

$$T_{i,j} = \begin{cases} E\left(l_k \right) & if\ (i,j) \in b_k \\ 0 & otherwise \end{cases} \tag{2}$$

where $E\left(l_k \right)$ is $l_k \to \mathbb{R}^{C_0}$, i.e., a mapping function that converts textual information into word vectors. 0 denotes a zero vector, which corresponds to a region with no text, and all pixels in each b_k share the same row-level word vector. The text embedding map T has the same spatial size and number of channels as image x. After obtaining the text embedding map T, in order to realize the fusion of multimodal features, it needs to be input to the text feature extraction module to extract text features. Since the convolutional neural network is good at learning deep features and can also map text features to the same subspace as visual features, it is used to extract multiscale text features from text embedding map Fig. A_i. where A_i has the same spatial size and number of channel dimensions as those extracted by the backbone network, the text feature extraction module is composed of four convolutional blocks, each of which contains convolutional layers and regularization layers. Where the size of the convolution kernel is 3 × 3, the step size is 2 and the edge padding is 1.

3.1.3 Feature Fusion Module

Multimodal perception mechanism has good feature extraction and learning ability, in this paper, we adopt ResNet as the backbone network to extract features and utilize it to map different modal spaces into a shared semantic subspace so as to deeply fuse multimodal features [18]. Visual information can easily recognize larger target regions, and textual information is important for distinguishing visually similar regions [19]. The textual features and visual features are added together, and the fused multimodal features are fed into the backbone network to extract multiscale features, and the textual features are incorporated many times in the extraction

process to enrich the feature information and realize the deep fusion of multimodal features. The feature fusion module is shown in Fig. 3, firstly, visual feature V_2 is extracted from the document image, and then textual feature A_2 is fused with it to obtain multimodal feature C_2. C_2 is input into the backbone network to obtain feature V_3, and fused with textual feature A_3 to obtain C_3, which can retain more information in the feature map by adding textual features. In this way, C_4 and C_5 are generated similarly, and the generation of feature C_i is defined as follows:

$$\begin{cases} C_i = V_i + A_i & i = 2,3,4,5 \\ V_2 = conv(x) & \\ V_{i+1} = convC_i & i = 2,3,4 \end{cases} \tag{3}$$

Where $conv$ is a convolutional layer, text features are defined as A_i and x represents the image. The information in the feature map is made richer by incorporating text features into the lower and higher layers.

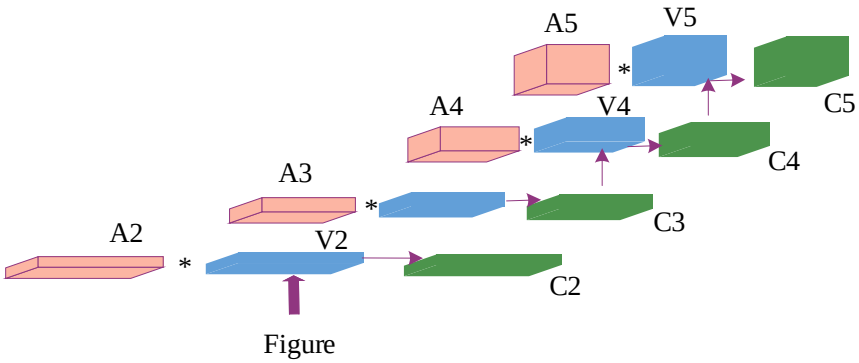


Figure 3 Feature fusion module

In the feature fusion module, different modal spaces can be mapped into a shared semantic subspace to fuse features from different modalities. Visual information contains higher level feature representations and textual information contains lower level feature representations, by fusing the complementary information of both, the fused feature information is made richer than the previous single modality.

3.2 Target Detection Accuracy Improvement Calculations

3.2.1 Introduction of attention mechanisms

In the process of feature fusion, features of different modalities are weighted and fused using an attention mechanism to highlight important features and suppress redundant information [20]. After obtaining the feature V_i of image and the feature A_i of text, in order to realize the interaction between the information of the two modalities and fully exploit the complementary ability of text features to image features, this paper proposes a

collaborative attention module. The computational process of this module is shown in Fig. 4, where the similarity between each image feature and all text features is obtained by calculating the attention weight A_x of each feature vector in V_i to all feature vectors in A_i. The greater the similarity, the closer the semantic information between that text feature and image feature. The similarity is then used as the weight to weight and sum the corresponding text features to obtain a new feature $\hat{A}_i$. $\hat{A}_i$ and V_i are spliced to integrate the text features into the image features to obtain the enhanced image feature $\hat{V}_i$, which realizes the enhancement of text features to image features.

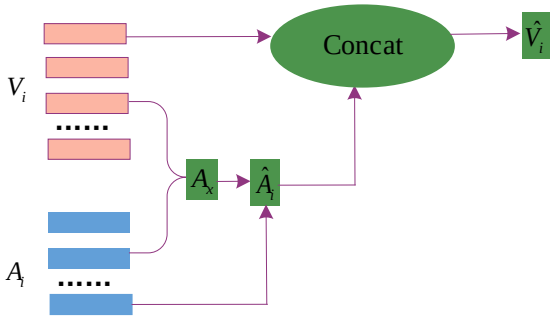


Figure 4 Collaborative attention module

The specific calculation process of the collaborative attention module proposed in this paper is as follows:

(1) The computation of attention weight matrices A_x, A_x is shown in Eq. (4) and Eq. (5):

$$S = V_i W \left(A_i\right)^T \tag{4}$$

$$A_x = soft\max\left(S\right) \in \mathbb{R}^{m \times n} \tag{5}$$

Where, $S \in \mathbb{R}^{m \times n}$ is the similarity matrix between V_i and A_i, $W \in \mathbb{R}^{C_2 \times C_3}$ is obtained by neural network training, and S_{ij} represents the similarity between the i th image feature vector and the j th text feature vector. The $soft\max$ operation is performed on S by row to obtain the attention weight matrix of image features to text features A_x. Where A_{ij}^x denotes the attention value of the i th image feature vector to the j th text feature vector.

(2) Compute the feature matrix $\hat{A}_i$.

Matrix multiply A_x and A_i to obtain the new feature matrix $\hat{A}_i$, $\hat{A}_i$ is calculated as shown in Equation (6):

$$\hat{A}_i = A_x A_i \tag{6}$$

where $\hat{V}_i \in \mathbb{R}^{m \times C_2}$.

(3) The computation of image features enhanced by text features $\hat{V}_i$, $\hat{V}_i$ is shown in Equation (7):

$$\hat{V}_i = Concat\left(V_i, \hat{V}_i\right) \tag{7}$$

where $\hat{V}_i \in \mathbb{R}^{m \times (C_2 \times C_3)}$, $\hat{V}_i$ have a total of m feature vectors, each of which contains information from both image and text modalities, for the detection of target objects in computer vision images.

3.2.2 Collecting contextual information

Contextual information in computer vision is information inferred from its surrounding pixels, and contextual information helps to improve the performance of each task. Therefore, this paper introduces a visual context module to extract visual context features, enrich the visible class context based on the dependencies existing in the former context, and also migrate it to the context of the invisible class. Utilizing contextual information to assist target detection can improve the model's ability to adapt to complex scenes.

To collect valid contextual information in the visual features, P is used as input and d is the channel dimension of the features, and the contextual features $\left\{h_j\right\}_{j=1}^{3} \in \mathbb{R}^{H \times W \times d}$ of the same resolution are first generated by 3 null convolution 2, $\left\{h_j\right\}_{j=1}^{3}$ containing more felt contextual information so that more valid contextual features can be selected. To adaptively select in $\left\{h_j\right\}_{j=1}^{3}$ based on the image context information, a convolutional layer is used to map the spliced $\left\{h_j\right\}_{j=1}^{3}$ into a weight matrix $A \in \mathbb{R}^{H \times W \times 3}$, where each $A_j \in \mathbb{R}^{H \times W}$ corresponds to the weight of h_j. Then A_j is multiplied with h_j to adjust the weight of context information of each pixel to get $M \in \mathbb{R}^{H \times W \times d}$, and the multi-scale context features are selected by this attention mechanism. To increase the feature diversity and improve the generalization performance of the model, $\sigma \in \mathbb{R}^{H \times W \times d}$ and $\mu \in \mathbb{R}^{H \times W \times d}$ are obtained by randomly sampling M using bilinear interpolation, and the sampled contextual features $N \in \mathbb{R}^{H \times W \times d}$ are obtained through Eq. (8):

$$N = \mu + \varepsilon\sigma \tag{8}$$

where ε is a random scalar obeying normal distribution generated from σ. The final contextual feature $Z \in \mathbb{R}^{H \times W \times d}$ is obtained by random sampling from N. For subsequent optimization of Z, Z is reshaped into visual contextual feature $F_v \in \mathbb{R}^{H \times W \times d}$ by reshaping operation to obtain Eq:

$$F_v = reshape(Z) \tag{9}$$

By introducing contextual information in the multimodal perception mechanism, the model can judge the reasonableness of the target appearance, avoid some unreasonable detection results, and improve the reliability and accuracy of detection [21].

4. Computer vision target detection accuracy verification

4.1 Experimental environment

In this paper, the effectiveness of multi-target perception mechanism is verified using the commonly used public dataset DIOR-R. The DIOR-R dataset contains a total of 23,463 computer images and 192,518 instance samples, which covers a large number of common target classes. The detection difficulties of this dataset include large target scale differences, arbitrary imaging direction, dense targets, complex background interference, and diverse shooting weather. The hardware platform configuration used for the experiments in this paper is shown in Table 1, and Pytorch deep learning development framework is used.

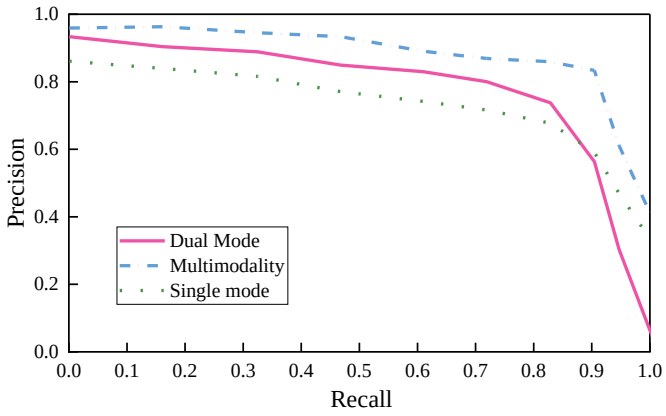
Table 1 Experimental environment and configuration

Category	Experimental configuration and environment
CPU	11thGenIntel(R)Core(TM)i7-11800H2.30GHz
GPU	NVIDIAGeForceRTX3060LaptopGPU
Operating System (OS)	Windows11
CUDA Version	CUDA11.1

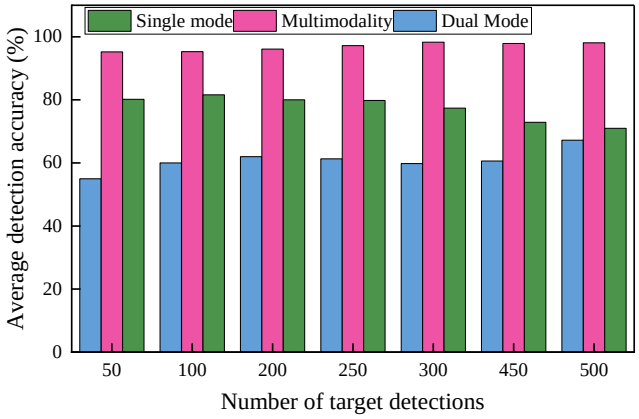
4.2 Comparison of target detection accuracy

4.2.1 Comparison of accuracy

In the verification of target detection accuracy, the use of average accuracy and checking rate for testing is an important indicator of model performance. Fig. 5 shows the results of the comparison of the check accuracy and average accuracy for different modalities, and the check accuracy and check all accuracy are shown in Fig. 5(a). It can be seen that compared with unimodal and bimodal, the multimodal perception mechanism has the best check-accuracy and check-full-rate, with the highest balanced position around 0.96. The accuracy improvement is better in real computer vision target detection. The average accuracy is shown in Fig. 5(b), where the average accuracy of the multimodal perception mechanism is above 95% for different detection targets, and the average accuracy of both unimodal and bimodal is not more than 85%. This further illustrates the superior performance of the multimodal perception mechanism, which can be applied to the study of detection accuracy enhancement in computer vision.



(a) Precision and recall curves



(b) Average accuracy comparison results

Figure 5 Comparison results of different modes

4.2.2 Interference testing

To further validate the performance of the multimodal perception mechanism in computer vision detection, the method in this paper is compared with 3D technology and 2D detection methods, and the average detection accuracy under interference conditions is shown in Figure 6. Taking partial occlusion and low-resolution interference conditions as examples, the average accuracies of the multimodal perception mechanism are 93.36 and 91.56, which are higher than the other two detection methods. Although 3D technology detection under contrast variation reaches 95.19, it is still lower than the 95.57 of multimodal perception.Under color interference conditions, the 2D detection technology has the worst performance of 63.11, and the multimodal perception still has the highest average accuracy of 93.28.It can be seen by the interference test that the multimodal perception machine, with feature fusion as well as the introduction of the attention mechanism, is able to adapt to the detection task better under different The multimodal perceptual machine with feature fusion and the introduction of the attention mechanism can better adapt to the detection task in different interference conditions, and has the best performance in terms of average detection accuracy compared with the other two methods, and has a stable anti-interference performance.

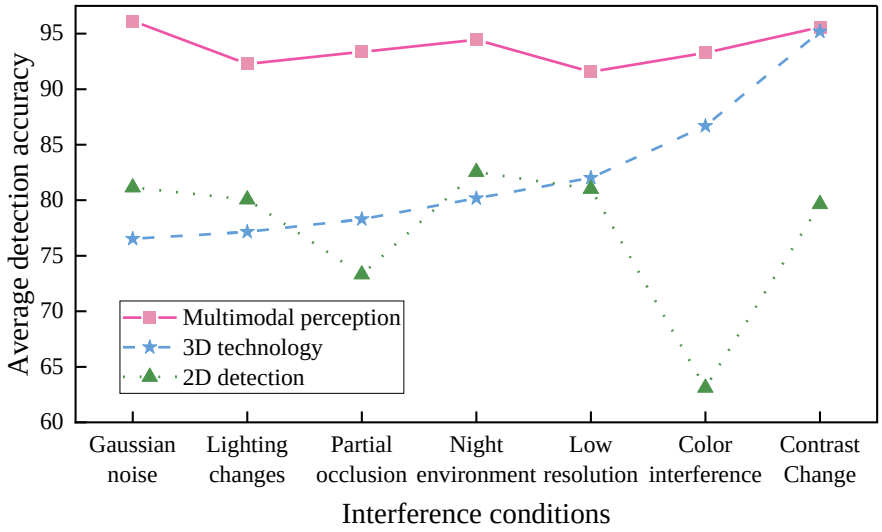


Figure 6 Average detection accuracy under interference conditions

4.3 Ablation experiments

In order to verify the effectiveness of each visual module in this paper's method, ablation experiments were conducted on the DI-OR-R dataset. The results of the ablation test are shown in Table 2, which shows that when the model does not use the attention mechanism and contextual information, its mAP is only 66.3%. This indicates that there is a large room for improvement in the performance of the underlying model when dealing with complex detection tasks. When only the attention mechanism is used, the mAP is 73.7%, which suggests that the attention mechanism is helpful in improving the model's performance by allowing the model to focus on key parts of the input data. When only contextual information is used, the mAP is 73.9%, which again shows the importance of contextual information for improving model performance because the multimodal perception mechanism provides additional information about the environment surrounding the input data. When both the attentional mechanism and contextual information are used, the mAP is significantly increased to 95.4%, which suggests that there is complementarity between the two techniques and that they can significantly improve the performance of target detection. These results suggest that the combined use of attentional mechanisms and contextual information is a very effective approach for target detection in computer vision.

Table 2 Ablation test results

Attention Mechanism	Contextual Information	mAP/%
×	×	66.3
√	×	73.7
×	√	73.9
√	√	95.4

In order to verify the detection ability of the model in real application scenarios, Table 3 shows the comparison results of the detection performance, which are analyzed as follows:

(1) The accuracy of the unimodal detection model is 63.27%, which means that among all the samples judged as positive samples by the model, only 63.27% are truly positive samples. The accuracy of the bimodal detection model is 72.21%, which is a significant improvement compared to the unimodal model, indicating that the bimodal model is more accurate in judging positive samples. The accuracy of the multimodal detection model is 96.64%, which is the highest among the three models, showing the strong ability of the multimodal model in accurately identifying positive samples.

(2) The recall of the unimodal detection model is 75.67%, indicating that the model is able to identify 75.67% of all true positive samples. The recall of the bimodal detection model is 84.14%, which is a large improvement compared to the unimodal model, indicating that the bimodal model is able to cover more positive samples. The recall of the multimodal detection model is 95.58%, which is almost close to 100%, indicating that the multimodal model is almost perfect in detecting positive samples.

(3) The AP_{50} of the unimodal detection model is 64.99%, indicating that the average precision of the model is about 65% at all thresholds when the IOU threshold is 0.5. The AP_{50} of the bimodal detection model is 75.37%, which is a significant improvement over the unimodal model. The AP_{50} of the multimodal detection model is 96.36%, which is much higher than the previous two, indicating that the multimodal model performs very well at an IOU threshold of 0.5.

(4) The $AP_{50:95}$ of the unimodal detection model is 50.50%, which is a comprehensive measure of the model's performance under multiple IOU thresholds. Lower values indicate that the model performs relatively poorly in response to different IOU thresholds. The $AP_{50:95}$ for the bimodal detection model is 52.54%, which is an improvement but not a significant one. The multimodal detection model has an $AP_{50:95}$ of 80.13%, which is the highest among the three models, indicating that the multimodal model has a more stable and excellent performance in dealing with different IOU thresholds.

Table 3 Detection performance on the DI-OR-R dataset

Detection model	Accuracy	Recall	AP_{50}	$AP_{50:95}$
Single modality	63.27%	75.67%	64.99%	50.50%
Dual modality	72.21%	84.14%	75.37%	52.54%
Multi-modality	96.64%	95.58%	96.36%	80.13%

The multimodal detection model shows extremely high levels of both precision, recall and AP metrics. This indicates that by combining information from multiple modalities, the model is able to better understand and recognize the target, thus improving the detection performance. In computer vision target detection, multimodal detection models are more suitable for complex and diverse scenes.

5. Conclusion

In this paper, we propose to apply the multimodal perception mechanism in the application of target detection accuracy enhancement in computer vision, which significantly improves the accuracy of target detection by introducing the attention mechanism and contextual information. The test results show that by comparing and

analyzing the performance of unimodal, bimodal and multimodal target recognition algorithms, the average accuracy of the multimodal perception mechanism is above 95%, and the average accuracy of unimodal and bimodal is not more than 85%. The performance comparison of different models on the DI-OR-R dataset shows that the multimodal detection model exhibits excellent performance on the DI-OR-R dataset, with a high precision of 96.64%, a recall of 95.58%, and a high AP_{50} and $AP_{50:95}$ of 96.36% and 80.13%, respectively. The average precision remains the highest under different interference conditions, outperforming 3D techniques and 2D detection methods. In conclusion, multimodal detection performs well in computer vision target detection, especially in complex scenes where the detection average accuracy remains above 90. Future research can further explore the application of multimodal perception mechanisms, combining more modal data and advanced techniques to improve the performance of computer vision target detection.

References

[1] Wagner, R., Matuschek, M., Knaack, P., Zwick, M., & Geiß, M. (2023). IndustrialEdgeML-End-to-end edge-based computer vision systemfor Industry 5.0. *Procedia Computer Science*, *217*, 594-603.

[2] Qian, R., Lai, X., & Li, X. (2022). 3D object detection for autonomous driving: A survey. *Pattern Recognition*, *130*, 108796.

[3] Bouraya, S., & Belangour, A. (2021). Deep learning based neck models for object detection: a review and a benchmarking study. *International Journal of Advanced Computer Science and Applications*, *12*(11).

[4] Yi, S., Zhang, G., Qian, C., Lu, Y., Zhong, H., & He, J. (2022). A multimodal classification architecture for the severity diagnosis of glaucoma based on deep learning. *Frontiers in neuroscience*, *16*, 939472.

[5] Al-Tameemi, I. K. S., Feizi-Derakhshi, M. R., Pashazadeh, S., & Asadpour, M. (2023). Interpretable multimodal sentiment classification using deep multi-view attentive network of image and text data. *IEEE Access*.

[6] Li, Z., Du, Y., Zhu, M., Zhou, S., & Zhang, L. (2022). A survey of 3D object detection algorithms for intelligent vehicles development. *Artificial Life and Robotics*, 1-8.

[7] Xu, C., Li, Z., Jiang, D., Yun, J., Liu, Y., Liu, Y., ... & Ying, S. (2021). 3D object detection based on synthetic RGB image. *International Journal of Wireless and Mobile Computing*, *20*(1), 70-76.

[8] Bengamra, S., Mzoughi, O., Bigand, A., & Zagrouba, E. (2024). A comprehensive survey on object detection in Visual Art: taxonomy and challenge. *Multimedia Tools and Applications*, *83*(5), 14637-14670.

[9] Cao, D., Chen, Z., & Gao, L. (2020). An improved object detection algorithm based on multi-scaled and deformable convolutional neural networks. *Human-centric Computing and Information Sciences*, *10*(1), 14.

[10] Tsai, T. H., & Yang, C. C. (2023). A real-time surveillance system with multi-object tracking. *Multidimensional Systems and Signal Processing*, *34*(4), 767-791.

[11] Duan, S., Lu, N., Lyu, Z., Liu, G., & Cao, B. (2022). An anchor box setting technique based on differences between categories for object detection. *International Journal of Intelligent Robotics and Applications*, *6*(1), 38-51.

[12] Muralidhara, S., Hashmi, K. A., Pagani, A., Liwicki, M., Stricker, D., & Afzal, M. Z. (2022). Attention-guided disentangled feature aggregation for video object detection. *Sensors*, *22*(21), 8583.

[13] Yang, W. J., Liow, W. J., Chen, S. F., Yang, J. F., Chung, P. C., & Mao, S. (2022). Improved vehicle detection systems with double-layer LSTM modules. *EURASIP Journal on Advances in Signal Processing*, *2022*(1), 7.

[14] Bairavel, S., & Krishnamurthy, M. (2020). Novel OGBEE-based feature selection and feature-level fusion with MLP

neural network for social media multimodal sentiment analysis. *Soft Computing*, *24*(24), 18431-18445.

[15] Guo, Y., Hu, T., Zhou, Y., Zhao, K., & Zhang, Z. (2022). Multi-channel data fusion and intelligent fault diagnosis based on deep learning. *Measurement Science and Technology*, *34*(1), 015115.

[16] Ge, T., Luo, X., Wang, Y., Sedlmair, M., Cheng, Z., Zhao, Y., ... & Chen, B. (2023). Optimally Ordered Orthogonal Neighbor Joining Trees for Hierarchical Cluster Analysis. *IEEE Transactions on Visualization and Computer Graphics*.

[17] Wang, H., & Li, F. (2022). A text classification method based on LSTM and graph attention network. *Connection Science*, *34*(1), 2466-2480.

[18] Wang, Y., Gu, Y., Yin, Y., Han, Y., Zhang, H., Wang, S., ... & Quan, D. (2023). Multimodal transformer augmented fusion for speech emotion recognition. *Frontiers in Neurorobotics*, *17*, 1181598.

[19] Wang, J., & Liu, S. (2022). Visual information computing and processing model based on artificial neural network. *Computational Intelligence and Neuroscience*, *2022*(1), 4713311.

[20] Wang, Z., Guo, J., Zeng, L., Zhang, C., & Wang, B. (2022). MLFFNet: Multilevel feature fusion network for object detection in sonar images. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-19.

[21] Wang, Y., Mao, K., Chen, T., Yin, Y., He, S., & Chen, G. (2023). Accelerating real-time object detection in high-resolution video surveillance. *Concurrency and Computation: Practice and Experience*, *35*(18), e6307.

ABOUT THE AUTHOR

Li Tao was born in Yinchuan, Ningxia. China, in 2000. He is currently studying at the School of Electronic and Electrical Engineering, Ningxia University. His main research focuses on image and signal processing, as well as artificial intelligence.

E-mail: albert2178@163.com