ISSN:2790-1688 Volume-15-(2025)

Modeling and Application of Deep Multimodal Integration in Intelligent Perception Systems

Minqi Peng *

The Chinese University of Hong Kong, CUHK, Hong Kong, China

menzelpeng@link.cuhk.edu.hk

Abstract. To address the limitations of single-modal perception in complex dynamic environments—such as insufficient robustness—and the challenges of multimodal fusion—including semantic alignment difficulties and high computational overhead—this paper proposes a dynamically weighted cross-modal Transformer framework (DW-CMTransformer). Firstly, a modal reliability evaluation module is constructed based on information entropy, which quantifies the confidence of heterogeneous sensors such as vision, lidar and voice in real time and generates adaptive weights. Secondly, the hierarchical attention mechanism is designed, and the two-way semantic alignment of multi-modal sequences is realized by using cross-modal Transformer at the feature level, and the single-modal output is weighted and fused at the decision level to ensure that the system performance remains stable when any mode fails. Finally, knowledge distillation is introduced, and the teacher model with 40M parameters is compressed to 5M, which increases the reasoning speed by 5 times and only loses 0.7% mAP. Experiments on nuScenes 3D target detection task show that the average mAP of DW-CMTransformer is 66.1%, which is 7.9% and 4.3% higher than Late Fusion and MMFN (Multimodal Fusion Network) baselines respectively, and the performance degradation is the smallest in the scene with single mode missing. This study provides an efficient and scalable new paradigm of multi-modal deep fusion for robust intelligent perception in edge computing environment, and the results can be transferred to key fields such as medical diagnosis and industrial detection.

Keywords: Multimodal Deep Fusion; Intelligent Perception System; DW-CMTransformer; Attention Mechanism.

1. Introduction

At the critical stage of AI's transition to cognitive intelligence, intelligent perception system, as the core carrier of human-machine-object ternary integration, is facing the inherent bottleneck of monomodal perception. The recognition rate of traditional vision system drops sharply in direct glare or rainy and foggy weather [1], and the word error rate (WER) of voice interaction module is high in factory noise environment [2], and the failure of a single sensor is more likely to cause paralysis of the whole perception system. Perception schemes that rely on a single mode have natural vulnerability in complex dynamic environment, and the collaborative processing ability of multi-source heterogeneous data has become the key competitiveness of the next generation intelligent system.

Multi-modal fusion technology provides a new paradigm for breaking through the limitation of single mode by integrating multi-dimensional information such as vision, hearing and touch. Early research mainly focused on shallow fusion methods, such as feature stitching and decision voting [3-4]. Although these methods can improve the robustness of the system, there are two major defects: First, the semantic correlation between modes is not considered, which leads to the accumulation of feature space alignment errors; Secondly, the static fusion strategy can not adapt to the dynamic change of modal quality. In recent years, the end-to-end fusion framework driven by deep learning has achieved cross-modal semantic interaction by introducing attention mechanism, and made breakthroughs in tasks such as video description generation, but its computational complexity has increased exponentially, making it difficult to deploy on edge devices with limited resources [5-6].

The current research faces three challenges: (1) There are orders of magnitude differences in sampling rate, data format and semantic granularity of different sensors; (2) The modal reliability changes with time in the real scene, which requires the fusion strategy to have real-time adjustment

ability; (3) Real-time systems such as autonomous driving require the model to complete multimodal reasoning within 100ms, but the existing deep fusion models generally have the problem of delay. These challenges have given birth to an urgent need for a new generation of multimodal fusion framework, which not only needs to realize deep feature interaction at semantic level, but also has the ability of dynamic weight distribution and lightweight deployment.

This study proposes the Dynamic Weighting Cross-Modal Transformer (DW-CMTransformer) framework, which overcomes existing limitations through three major innovations: (1) Design a modal reliability evaluation module, and dynamically calculate the contribution of each mode based on the information entropy theory; (2) Construct a hierarchical attention mechanism to achieve two-way semantic alignment at the feature level and the decision level; (3) By introducing knowledge distillation technology, the large-scale fusion model can still maintain high accuracy after being compressed by 8 times. Experimental results show that this method can greatly improve the mAP of the baseline model on the nuScenes data set of the autonomous driving scene, and can still maintain a high recognition rate when the single mode fails. This study provides a new idea for robust perception in complex scenes, and its dynamic fusion mechanism can be extended to medical diagnosis, industrial detection and other fields.

2. Method design

2.1 Overview of the overall framework

Transformer model is a deep learning model based on self-attention mechanism, which is the core innovation of Transformer. It allows the model to pay attention to the information of all other words in the input sequence when processing a word, rather than relying only on the local information of a fixed window.By calculating the similarity between each input word and the representation of all other words, the weight of each input word in the current word representation is adjusted, so as to obtain the contextual understanding of the word [7]. In order to improve the model's ability to capture different semantic relationships, Transformer adopts multi-head attention mechanism. By dividing the self-attention mechanism into multiple heads, each head can pay attention to different parts or different types of relationships in the sequence, thus getting more contextual information [8].

Because Transformer itself does not have the ability to deal with the sequence order, it is necessary to introduce position coding to preserve the position information in the sequence. Position coding is usually generated by sine and cosine functions, which is periodic, so that the model can learn the relative relationship between different word positions [9]. Its structure is shown in Figure 1, and the encoder is responsible for converting the input sequence into an intermediate representation. Each encoder layer consists of two main sublayers: a multi-head self-attention mechanism layer and a fully connected feedforward neural network layer. Residual connection and layer normalization are adopted between these two sub-layers. The decoder generates the target sequence according to the output of the encoder. Similar to the encoder, the decoder is also composed of several identical layers, but each decoder layer contains an additional encoder-decoder attention layer to pay attention to the output of the encoder in addition to the multi-head self-attention mechanism layer and the feedforward neural network layer.

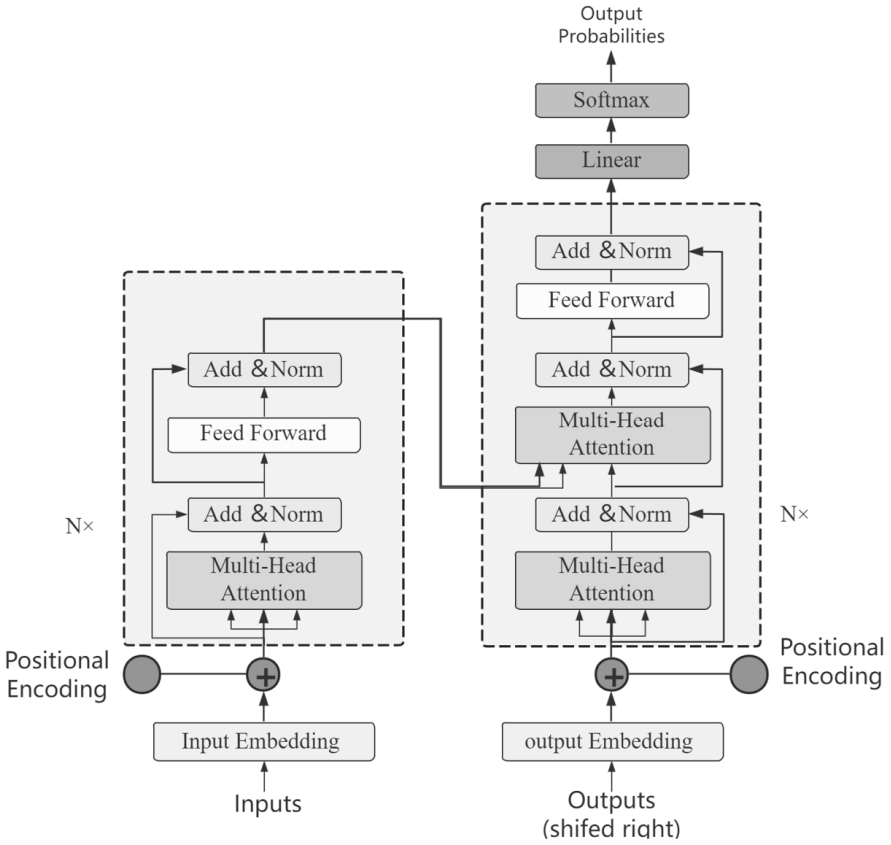

Figure 1 Transformer model structure

The framework of DW-CMTransformer proposed in this paper aims to solve the vulnerability problem of monomodal perception in complex environment [10-11]. This method realizes the deep fusion and efficient deployment of multimodal data through modal reliability evaluation, hierarchical attention mechanism and knowledge distillation technology. The DW-CMTransformer framework consists of three core modules, as shown in Figure 2.

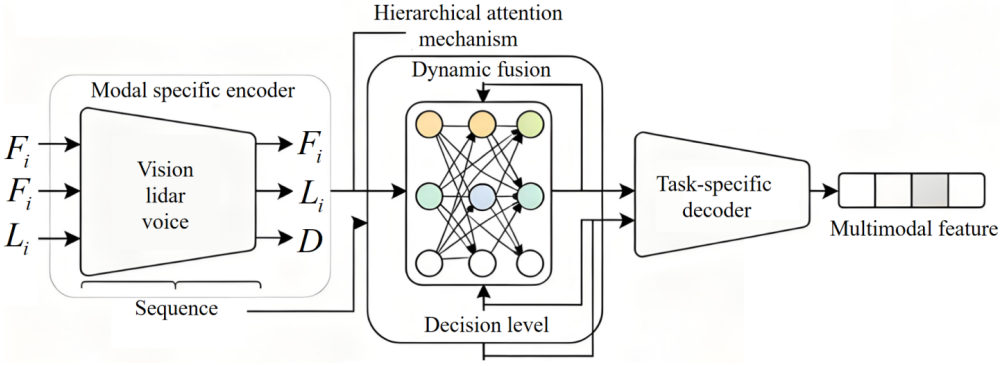

Figure 2 DW-CMTransformer

For mode-specific encoders, each mode (vision, lidar, voice) uses an independent encoder to extract feature representation. Let the number of input modes be M and the feature of mode i be $F_i \in R^{L_i \times D}$, where L_i is the sequence length and D is the feature dimension. The dynamic fusion module includes modal reliability evaluation and hierarchical attention mechanism to realize the fusion of feature level and decision level. The task-specific decoder outputs perceptual results based on the fusion features.

2.2 Modal reliability evaluation

Based on the information entropy theory, the contribution weight of each mode is calculated dynamically. Entropy measures the uncertainty of modal output: the higher the entropy, the lower the reliability [12]. For each mode i, the class probability distribution $p_i \in R^C$ is extracted from its encoder output, where C is the number of classes. p_i is obtained by softmax function, which indicates the prediction confidence of mode$i$.

Calculate the information entropy H_i of mode i:

$$H_i = -\sum_{c=1}^{C\sum_i(c)} p_i(c) log \tag{1}$$

Where $p_i(c)$ is the prediction probability of modal i for class c. The value range of H_i is$[0, log\ C]$, and the larger the value, the higher the uncertainty.

Define the reliability score r_i as the complement of entropy;

$$r_i = 1 - \frac{H_i}{log\ C} \tag{2}$$

Normalized to [0,1] interval. But for simplicity, you can directly use:

$$r_i = 1 - H_i \tag{3}$$

Then normalize the weights by softmax:

$$w_i = \frac{exp(x_i)}{\sum_{j=1}^{M} exp(x_j)} \tag{4}$$

Where w_i is the dynamic weight of modal i and M is the total number of modes. The weight w_i is dynamically adjusted with the input data. For example, in rainy and foggy weather, the visual mode entropy increases and the weight decreases, while the weight of lidar increases.

2.3 Hierarchical attention mechanism

Realize the two-way semantic alignment between feature level and decision level, and ensure the complementary enhancement between modes [13]. Cross-modal interaction is carried out at the feature level, and a cross-modal Transformer encoder is used [14]. Firstly, the modal characteristics are preliminarily weighted by using the reliability weight w_i:

$$\hat{F}_i = w_i F_i \tag{5}$$

Then, all weighted features are spliced into $F = \left[\hat{F}_1; \hat{F}_2; \cdots; \hat{F}_M\right]$, where,

$$L = \sum_{i=1}^{M} L_i \tag{6}$$

Next, feature fusion is achieved using Multi-Head Self-Attention (MSA):

$$Z = LayerNorm\big(F + MSA(F)\big) \tag{7}$$

Where MSA is calculated as:

$$MSA(F) = Concat(head_1, \cdots, head_h)W^O$$
$$head_k = Attention\big(FW_k^Q, FW_k^K, FW_k^V\big) \tag{8}$$

The formula of attention mechanism is:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{9}$$

Here, Q, K, V is the matrix of query, key and value, which is obtained by F projection; FW_k^Q, FW_k^K, FW_k^V is a learnable weight); d_k is the dimension of the key, which is used for scaling; h is the head number. Through multi-layer Transformer block, feature-level bidirectional alignment is realized (each modal feature can pay attention to other modes).

At the decision-making level, the independent output of each mode is integrated to enhance robustness [15]. Each mode i generates a decision output $d_i \in R^C$ through a task header. Then, the reliability weight w_i is used for weighted fusion:

$$d = \sum_{i=1}^{M} w_i d_i \tag{10}$$

The final decision is made by d through softmax to get the probability output. This hierarchical design ensures that decision-level fusion can maintain performance even if a single mode fails.

2.4 Knowledge distillation technology

In order to compress the model, knowledge distillation is introduced. The complete DW-CMTransformer is used as the teacher model, and the compressed lightweight version is used as the student model. Distillation loss function combines task loss and distillation loss;

$$L_{total} = \alpha L_{task} + (1-\alpha)T^2 KL(\sigma(z_s/T)\|\sigma(z_t/T)) \quad (11)$$

Where L_{task} is the task loss, which is used for the task output of the student model. z_s, z_t is the logits output of student and teacher models respectively (before decision-making level integration). T is a temperature parameter (usually $T > 1$), which is used to smooth the softmax output $\sigma(\cdot)$. $KL(\cdot\|\cdot)$ is KL divergence, which measures the difference of output between students and teachers. α is the balance weight. Through distillation, the parameters of the student model can be reduced by 8 times, while maintaining high accuracy.

3. Experimental verification

3.1 Experimental setup

Use the authoritative multimodal data set nuScenes in the field of autonomous driving. The data set contains 1000 driving scenes, providing images from six cameras, a point cloud of a 32-line laser radar, millimeter-wave radar data and other modal information. This experiment focuses on 3D object detection tasks. The visual mode uses ResNet-50+FPN as the encoder, and the point cloud mode uses PointPillar as the encoder. The number of layers of the Transformer encoder is 3, and the number of attention heads is 8. AdamW optimizer is used in training, and the initial learning rate is 0.0001.

3.2 Overall performance comparison

Overall performance comparison On the nuScenes verification set, the 3D object detection mAP pairs of different methods are shown in Table 1 below.

Table 1 Overall performance comparison (nuScenes verification set, mAP/%)

Model	Automobile	Pedestrian	Bicycle	Bus	Average mAP
Late Fusion	68.3	55.1	25.7	52.9	58.2
MMFN (Multimodal Fusion Network)	72.5	58.9	28.4	56.3	61.8
DW-CMTransformer（Ours）	76.8	63.5	32.1	60.2	66.1

As can be seen from Table 1, the DW-CMTransformer method proposed in this paper achieves the best performance in all object categories, with an average mAP of 66.1%, which is 7.9% and 4.3% higher than Late Fusion and MMFN baselines respectively. This fully proves the advantages of dynamic fusion mechanism and hierarchical attention in integrating multimodal information, which can capture the complementary semantics between modalities more effectively, thus improving the perception accuracy.

3.3 Robustness analysis under modal failure

In order to simulate the real-world sensor fault, a test subset is constructed, in which one or more modal data are randomly discarded. Figure 3 below shows the performance retention rate in two extreme cases of "missing visual mode" and "missing point cloud mode" (based on the mAP of the model with complete data as 100%).

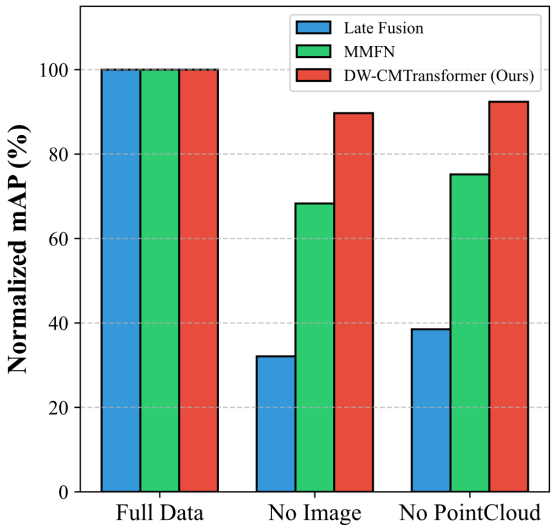

Figure 3 Robustness analysis histogram

Under the condition of "no image", Late Fusion performance plummeted (due to the complete failure of visual branches), MMFN decreased significantly, and our method had the smallest decline. Under the condition of "no point cloud", the trend is similar. Our model curve is the most gentle. The experimental results show that DW-CMTransformer shows excellent robustness when a single mode fails. This is mainly due to the dynamic weight allocation module. When a certain mode is missing, its information entropy will increase sharply, which will lead to the automatic reduction of reliability weight w_i, and the system will naturally reduce its dependence on the failed mode when fusing, and at the same time enhance its trust in the available modes, thus maintaining the stable output of the whole system.

3.4 Ablation experiment

The ablation experiment is designed to verify the contribution of each core component in the framework, and the results are shown in Table 2 below.

Table 2 Ablation experiment (average mAP/%)

Model version	Dynamic weight	Stratified attention	mAP
Baseline (Late Fusion)	-	-	58.2
A:+characteristic attention	-	✓	62.5
B: A+fixed weight fusion	-	✓	63.3
C: DW-CMTransformer (complete model)	✓	✓	66.1

Compared with Baseline, version A is obviously improved, which proves the effectiveness of cross-modal feature interaction. Version b uses fixed weight fusion on the basis of a, and its performance is slightly improved, but it is not as good as version C. The complete DW-CMTransformer (version C) achieves the highest performance, which clearly shows that the dynamic weight allocation mechanism and the hierarchical attention mechanism are complementary and indispensable, and they work together to maximize the performance.

3.5 Knowledge distillation effect

After being compressed by knowledge distillation technology, the parameter quantity of the student model is only 1/8 of that of the teacher model (about 5M parameters). As shown in Table 3, the reasoning speed of the student model is improved by nearly five times on the premise of minimal precision loss. The results show that the knowledge distillation technology successfully compresses the large-scale fusion model into a lightweight model, which greatly improves the reasoning efficiency with only 0.7% mAP loss, making it more suitable for edge computing scenarios with strict real-time requirements.

Table 3 Comparison of knowledge distillation effects

Model	Parameter quantity (M)	mAP (%)	Inference time (ms/ frame)
Teacher model (DW-CMTransformer)	40.1	66.1	120
Student model (after distillation)	5.0	65.4	25

The results show that this method not only performs best under ideal conditions, but also shows strong robustness in the real challenge of modal failure. At the same time, it ensures the feasibility of practical application through knowledge distillation, which provides a reliable technical path for intelligent perception in complex scenes.

4. Conclusion

Multi-modal deep fusion shows remarkable potential and advantages in intelligent perception system, especially in complex dynamic environment. The DW-CMTransformer with dynamic weight distribution proposed in this paper breaks through the limitations of the existing methods through three innovations. The modal reliability evaluation module dynamically calculates the contribution of each mode based on the information entropy theory, the hierarchical attention mechanism realizes the two-way semantic alignment between the feature level and the decision level, and the knowledge distillation technology keeps high accuracy after compressing the large-scale fusion model by 8 times. The experimental results show that DW-CMTransformer achieves the best performance on the nuScenes data set of autonomous driving scene, with an average mAP of 66.1%, which is 7.9% and 4.3% higher than Late Fusion and MMFN baselines respectively. In addition, when a single mode fails, the model shows excellent robustness and its performance decreases the least. Ablation experiment further verified the complementary effect of dynamic weight distribution mechanism and stratified attention mechanism. The knowledge distillation technology successfully compresses the large-scale fusion model into a lightweight model, which greatly improves the reasoning efficiency with only 0.7% mAP loss, making it more suitable for edge computing scenarios with strict real-time requirements. DW-CMTransformer not only performs best under ideal conditions, but also shows strong robustness in the real challenge of modal failure. At the same time, it ensures the feasibility of practical application through knowledge distillation and provides a reliable technical path for intelligent perception in complex scenes.

References

[1] Ye L ,Shiyang M ,Hongzhang W , et al.Deep learning based object detection from multi-modal sensors: an overview[J].Multimedia Tools and Applications,2023,83(7):19841-19870.

[2] Lin H X ,Xing D X ,Ren Z Y , et al.Prediction model of permeability index for blast furnace based on WD-NL-transformer[J].Ironmaking & Steelmaking,2025,52(9):1046-1057.

[3] Martínez H ,Catalán S ,Castelló A , et al.Characterization of quantized inference with transformer encoders on low power CPUs[J].The International Journal of High Performance Computing Applications,2025,39(6):803-821.

[4] Gao H ,Su L ,Zheng Y , et al.Load forecasting method based on Transformer multi-model fusion[J].Journal of Computational Methods in Sciences and Engineering,2025,25(6):5086-5097.

[5] Chen H ,Wang X X ,Zhang S R , et al.[The application and challenges of multi-modal data fusion based on deep learning in pathology] .[J].Zhonghua bing li xue za zhi = Chinese journal of pathology,2025,54(10):1032-1038.

[6] Tang J .Multi-modal fusion and transferable deep learning for rare disease detection: a CNN-Transformer framework with cross-domain adaptation on limited CT and MRI data[J].Advances in Engineering Innovation,2025,16(6):144-148.

[7] Peng Y ,Zheng Y ,Han L , et al.SIM-Net: A specific information-based multi-modal network for accurate diagnosis of breast lesions in multi-parametric MRI: A multi-center study.[J].Asian journal of surgery,2024,48(2):1284-1286.

[8] Wu C ,Wang Y ,Qiu S , et al.A bimodal deep learning network based on CNN for fine motor imagery[J].Cognitive Neurodynamics,2024,18(6):1-14.

[9] Gagan V ,Kumar A N ,Singh G T .Optimized vision transformer encoder with cnn for automatic psoriasis disease detection[J].Multimedia Tools and Applications,2023,83(21):59597-59616.

[10] Xiang G ,Sining W ,Ying Z , et al.Lightweight image super-resolution via multi-branch aware CNN and efficient transformer[J].Neural Computing and Applications,2023,36(10):5285-5303.

[11] Minming G ,Zhixiang C ,Kaiyu C , et al.RMPCT-Net: a multi-channel parallel CNN and transformer network model applied to HAR using FMCW radar[J].Signal, Image and Video Processing,2023,18(3):2219-2229.

[12] Huafeng L ,Junyu L ,Yafei Z , et al.A Deep Learning Framework for Infrared and Visible Image Fusion Without Strict Registration[J].International Journal of Computer Vision,2023,132(5):1625-1644.

[13] Xianwen D ,Jiacheng L ,Ke N , et al.Multiscale deep feature selection fusion network for referring image segmentation[J].Multimedia Tools and Applications,2023,83(12):36287-36305.

[14] Guanru T ,Teng Z ,Boyu H , et al.A noise-immune and attention-based multi-modal framework for short-term traffic flow forecasting[J].Soft Computing,2023,28(6):4775-4790.

[15] R. M B ,A. K R R .Optimal Score Level Fusion for Multi-Modal Biometric System with Optimised Deep Ensemble Technique[J].Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization,2023,11(5):1906-1920.