KNN adaptive dual attention for object detection

Hongrun Zhu[1,2], Zengmin Xu[1,2,3,*], Ruxing Meng[3], Longfei Liu[1,2]

1 School of Mathematics and Computing Science, Guangxi Colleges and Universities Key Laboratory of Data Analysis and Computation, Guilin University of Electronic Technology, Guangxi, Guilin 541004, China

2 Center for Applied Mathematics of Guangxi (GUET), Guangxi, Guilin 541002, China

3 Anview.ai, Guangxi, Guilin 541010, China

* Corresponding author(s). E-mail(s): xzm@guet.edu.cn

In construction site scenes, the key to detecting workers wearing safety helmets is to accurately locate the target position in the complex background and rapidly extract fine-grained feature information of the safety helmet for identification. Attention mechanisms assist deep models in focusing on salient features while suppressing irrelevant elements. However, most existing mixed attention mechanisms are not only complex in design but also fail to capture global contextual information. Although existing self-attention mechanisms can model long-range dependencies, they are computationally complex and ignore the expressions of local visual saliency. In this paper, we propose a lightweight KNN Adaptive Dual Attention (KADA) module. KADA employs a Gaussian Mixture Model to compute self-attention maps, which can reduce information loss by adaptively selecting the number of Gaussian components. Then, a KNN-based channel attention mechanism is embedded in self-attention computations, allowing adaptive selection of K channels for interaction within a low-dimensional manifold. This integration enables the capture of both global spatial information and fine-grained semantic information. Experimental results on the Safety Helmet Wearing Detection Dataset(SHWD) and the Complex Real-world Construction Site (CRCS) dataset demonstrate that KADA achieves a better balance between performance and computational complexity. Compared to various attention mechanisms, KADA achieves state-of-the-art results.

Additional Keywords and Phrases: Attention mechanism, Self attention, Channel attention, Helmet wearing detection

INTRODUCTION

In recent years, computer vision-based construction site monitoring and early warning systems have become important means for the digital transformation of the construction industry. Among them, incorporating visual intelligent technologies into safety helmet detection has been a hot research topic. Kelm [1]proposed a sensor-based approach for safety helmet detection, but it suffers from high costs and maintenance difficulties. Shrestha [2]used traditional image recognition algorithms for safety helmet detection, but these algorithms are slow and have low accuracy, making them unsuitable for real-time monitoring. With the development of deep learning, researchers have applied convolutional neural networks to safet[1]y helmet wearing detection [3,4]. However, in complex construction site scenes, different categories of safety helmet images exhibit small inter-class differences and large intra-class variations, making it challenging for deep models to distinguish fine-grained semantic characteristics, thereby affecting classification performance[5]. Moreover, high-level deep features have overly complex classification boundaries in high-dimensional space, making it difficult for visual models to capture rich contextual dependencies, which in turn affects localization performance. In common object detection systems, increasing the receptive field is achieved by stacking pooling and convolutional layers. However, repeated use of convolution and pooling operations with strides results in significant loss of spatial information. Attention mechanisms, such as self-attention Non-Local [6], EMANet [7], channel attention SENet [8], ECANet [9], and hybrid attention CBAM [10], DANet [11], have been widely employed to describe where and what to focus on by computing the interrelationships among data. Attention mechanisms have been extensively applied in computer vision. Non-Local Networks model each pixel to allow the fusion of features from all other positions, enabling the representation of features from different locations. However, the Non-Local module requires the generation of a huge attention map. CCNet aggregates the contextual information from all horizontal and vertical pixels along its horizontal and vertical paths for each pixel, resulting in complete image correlation captured by the attention module through iterative operations, while reducing computational complexity. EMANet does not compute the attention map for all pixels but constructs a set of bases and calculates attention scores on this basis to obtain contextual dependencies. However, these self-attention mechanisms only capture global spatial information and overlook fine-grained semantic representation. Each channel of the feature map is responsible for extracting different types of features [12]. SENet calibrates the importance of channels through squeeze and excitation modules, enhancing the network's expressive power by modeling interdependencies between channels. Sun [13]improved the accuracy of channel weight allocation through a multi-level feature interaction approach. Dan [14]introduced multi-stage channel attention for image dehazing in CycleGAN [15]. ECANet, utilize channel-wise interactions to capture feature correlations. However, these channel attention modules overlook spatial information and fail to capture contextual dependencies. CBAM adds a spatial attention module to the SE attention module, obtaining the dependencies between feature channels and feature spaces by serially applying attention along the

This research was supported by the National Natural Science Foundation of China (61862015), the Guangxi Natural Science Foundation (2024GXNSFAA010493), the Science and Technology Project of Guangxi (AD23023002), the Guangxi Key Research and Development Program (AB17195025).

channel and spatial axes. DANet computes channel attention and spatial attention in parallel. However, CBAM, DANet, and other hybrid attention mechanisms lack global dependencies, which are crucial for understanding the relative positional relationships of objects. Moreover, they compute attention scores separately in two independent modules, resulting in a large number of parameters and making them unsuitable for embedding in low-computational-power devices. CANet [16]aggregates spatial position information in both horizontal and vertical directions and embeds it into channel information, reducing parameter volume. However, from a spatial perspective, channel attention is applied globally [10], while CA only includes long-range dependencies between pixels in the spatial domain, lacking local dependencies.

The aforementioned methods have conducted research on attention mechanisms at different levels, but they still face the following problems:

1. Most self-attention mechanisms only include global spatial feature information and overlook fine-grained semantic feature expression at a granular level, which is crucial for fine-grained recognition.
2. Although most channel attention mechanisms possess the capability to capture fine-grained feature expression, they disregard spatial feature information, which is vital for utilizing spatial positional relationships in complex scenes for object detection.
3. While some hybrid attention mechanisms incorporate both spatial and channel feature information, they are divided into separate modules to calculate spatial and channel attention scores, resulting in high computational and parameter complexity. Consequently, they cannot be embedded in low-computational-power devices. Although some attention methods can reduce model parameters, they only provide global spatial dependencies and lack local dependencies, which are essential for understanding spatial features.

This paper proposes a plug-and-play scalable KNN adaptive dual attention mechanism, which incorporates global spatial feature information, local spatial information, and fine-grained semantic information in a balanced manner between parameter volume and detection performance. The contributions of this paper can be summarized as follows:

1. Introducing a Gaussian Mixture Models Adaptive self-attention mechanism (GMMA) that calculates the global attention map by adaptively selecting the number of Gaussian components, thereby obtaining global spatial feature information
2. A KNN Dual Attention Module (KADA) is proposed in this paper. Specifically, a KNN-based channel attention mechanism is introduced and embedded within the proposed Gaussian Mixture Model (GMM)-based adaptive self-attention module. This integration enables the module to capture local spatial information and fine-grained semantic information.
3. Integrating KADA into the YOLOv5 model and validating its feasibility and effectiveness through several datasets.

1 METHOD

1.1 Gaussian Mixture Models Adaptive self-attention

In construction site environments, the visual confusion between safety helmets worn by workers and background clutter poses a significant challenge. The task of object detection can be regarded as finding a mapping of data points from a high-dimensional feature space to a low-dimensional manifold, thereby capturing essential semantic information in a low-noise space. Within self-attention mechanisms represented by Non-Local, the computation of self-attention can be expressed as follows:

$$\mathbf{y}_i = \frac{1}{C(\mathbf{x})}\sum_{\forall j} f\left(\mathbf{x}_i, \mathbf{x}_j\right) g\left(\mathbf{x}_j\right) \tag{1}$$

where $f(\cdot)$ can be viewed as the mapping function, $\mathbf{y}_i$ represents the reconstruction of $\mathbf{x}_i$ by taking the weighted average of $g(\mathbf{x}_j)$ using weights $\frac{1}{C(\mathbf{x})} f\left(\mathbf{x}_i, \mathbf{x}_j\right)$ to obtain the correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$. The functions $f(\cdot)$ and $g(\cdot)$ are typically implemented using the Softmax function, which can be expressed as:

$$\mathbf{y} = softmax\left(\mathbf{x}^T W_\theta^T W_\phi \mathbf{x}\right)\left(W_\sigma \mathbf{x}\right)^T = softmax\left(\theta(\mathbf{x})^T \phi(\mathbf{x})\right)\sigma(\mathbf{x}) \tag{2}$$

where $\theta(\mathbf{x})^T \in \mathbb{R}^{N\times C}$, $\phi(\mathbf{x}) \in \mathbb{R}^{C\times N}$, $\sigma(\mathbf{x})^T \in \mathbb{R}^{N\times C}$. The above equation can be represented as the reconstruction of $\mathbf{x}_i$ using a set of bases. However, this set of bases is excessively complete (as $\mathbf{x}_j$ represents all the pixels in the image), leading to a significant amount of redundant information. Therefore, when reconstructing $\mathbf{x}_i$, it is only necessary to find a set of bases in the high-dimensional feature space that satisfies the following criteria: (1) the reconstructed feature map can adequately restore the relevant information from the original feature map, (2) the reconstructed features possess low-rank characteristics, indicating that they lie within a low-dimensional manifold in the high-dimensional space, and (3) the reconstructed feature map can be partitioned into multiple clusters, with small intra-cluster differences and large inter-cluster differences.

In the Gaussian Mixture Model (GMM), different Gaussian components can model the features within the low-dimensional manifold. Therefore, finding this set of components is equivalent to finding the set of bases. When solving for the GMM component parameters, the EM algorithm is commonly employed for iterative estimation. Specifically, considering an input feature map $\mathbf{x} \in \mathbb{R}^{N \times C}$, where $N = H \times W$ and $\mathbf{x}_i \in \mathbb{R}^C, i = 0,1,...,N$ represent the i pixel's c channel in the feature map, the bases are initialized as μ. Here, $\mu \in \mathbb{R}^{K \times C}$ and the latent variable $\mathbf{z}$ represent the K component of the GMM, where $\mathbf{z} \in \mathbb{R}^{N \times K}$. Consequently, the likelihood function of the Gaussian Mixture Model can be derived as:

$$p(\mathbf{x}, \mathbf{z} \mid \mu, \mathbf{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{\mathbf{z}_{nk}} \mathcal{N}(\mathbf{x}_n \mid \mu_k, \mathbf{\Sigma}_k)^{\mathbf{z}_{nk}} \tag{3}$$

where π_k is the weighted sum of multiple Gaussian components, and $\boldsymbol{\mu}_k, \mathbf{\Sigma}_k$ represents the parameters of the k component. $\gamma(\mathbf{z}_{nk})$ denotes the responsibility of the k component for the n pixel. According to the relationship between joint probability density and marginal probability density, we have:

$$p(\mathbf{z} \mid \mathbf{x}, \mu, \mathbf{\Sigma}, \boldsymbol{\pi}) = \frac{\prod_{n=1}^{N} \prod_{k=1}^{K} \left[\pi_k \mathcal{N}(\mathbf{x}_n \mid \mu_k, \mathbf{\Sigma}_k) \right]^{\mathbf{z}_{nk}}}{\prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n, \mu_k, \mathbf{\Sigma}_k)} \tag{4}$$

where $\prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n, \mu_k, \mathbf{\Sigma}_k)$ is a constant. Combining equation (3) and equation (4), we can obtain the marginal distribution of $\mathbf{z}$ as follows:

$$p(\mathbf{z} \mid \mathbf{x}, \mu, \mathbf{\Sigma}, \boldsymbol{\pi}) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\pi_k \mathcal{N}(\mathbf{x}_n \mid \mu_k, \mathbf{\Sigma}_k) \right]^{\mathbf{z}_{nk}} \tag{5}$$

In the E-step, the expectation of $\mathbf{z}_{nk}$ is computed as follows:

$$\begin{aligned} \mathbb{E}[\mathbf{z}_{nk}] &= \frac{\sum_{\mathbf{z}_n} \mathbf{z}_{nk} \prod_{k'} \left[\pi_{k'} \mathcal{N}(\mathbf{x}_n \mid \mu_{k'}, \mathbf{\Sigma}_{k'}) \right]^{\mathbf{z}_{nk'}}}{\sum_{\mathbf{z}_n} \prod_{j} \left[\pi_j \mathcal{N}(\mathbf{x}_n \mid \mu_j, \mathbf{\Sigma}_j) \right]^{\mathbf{z}_{nj}}} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \mu_k, \mathbf{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n \mid \mu_j, \mathbf{\Sigma}_j)} \\ &= \gamma(\mathbf{z}_{nk}) \end{aligned} \tag{6}$$

The above equation can be expressed in a general form as follows:

$$\gamma(\mathbf{z}_{nk}) = \frac{\pi_k \mathcal{K}(\mathbf{x}_n \mid \mu_k, \mathbf{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{K}(\mathbf{x}_n \mid \mu_j, \mathbf{\Sigma}_j)} \tag{7}$$

In visual attention mechanisms, $\mathcal{K}(a,b)$ in the above equation is typically represented by $softmax(\cdot)$. After completing the E-step, the M-step is then computed by taking the partial derivatives of μ_k and $\mathbf{\Sigma}_k$ in equation (3), resulting in:

$$\begin{aligned} \mathbf{\Sigma}_k^{new} &= \frac{1}{N_k} \pi_k \sum_{n=1}^{N} \gamma(\mathbf{z}_{nk}) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^{\mathrm{T}} \\ \mu_k^{new} &= \frac{1}{N_k} \pi_k \sum_{n=1}^{N} \gamma(\mathbf{z}_{nk}) \mathbf{x}_n, \text{ Here } N_k = \sum_{n=1}^{N} \gamma(\mathbf{z}_{nk}) \end{aligned} \tag{8}$$

After each iteration, the reconstructed $\mathbf{x}_n^{new}$ can be obtained as follows:

$$\mathbf{x}_n^{new} = \sum_{k=1}^{K} \pi_k \gamma(\mathbf{z}_{nk})^{new} \mu_k^{new} \tag{9}$$

After the convergence of GMM, we obtain the final $\mu^{(T)}$ and $\gamma(\mathbf{z})^{(T)}$, which can be used to reconstruct the feature map $\mathbf{x}$ as follows:

$$\hat{\mathbf{x}} = \mathbf{z}^{(T)} \mu^{(T)} \tag{10}$$

In the EMANet, a semantic segmentation model proposed by Li et al., the EMA unit is used to generate attention maps. The dimension K of the bases μ in the EMA unit needs to be specified during the initialization phase. However, in convolutional neural networks, there can be significant variations in the number of channels across different layers. For example, in YOLO v5, the number of channels in each module of the Backbone, Neck, and Head parts is quite different.

The information density varies across different-dimensional feature layers, and if they are uniformly projected into a fixed-dimensional space of K, important feature information from higher-dimensional spaces may be lost. Therefore, the dimension of the mapping space is dynamically adjusted according to $d = C / \gamma$, enabling the retention of sufficient original information while minimizing the dimensionality as much as possible.

For the above theoretical analysis, the pseudocode is as follows:

ALGORITHM 1: Adaptive Self-attention Mechanism Based on GMM

Input: feature map $\mathbf{x}$, latent variable $\mathbf{z}$

Compute the number of GMM components K, randomly initialize the parameters $\theta=(\pi_K,\boldsymbol{\mu}_K,\boldsymbol{\Sigma}_K)$, and denote them as θ^{old}.

for $epoch=1,...,N$, do

for $t=1,...,T$, do

compute $\mathbf{z}^{(t)}$, obtain $p\left(\mathbf{z}^{(t)}\mid\mathbf{x},\theta^{old}\right)$

optimize $\mathcal{Q}\left(\theta,\theta^{old}\right)=\sum_{\mathbf{z}}p\left(\mathbf{z}^{(t)}\mid\mathbf{x},\theta^{old}\right)\ln p(\mathbf{x},\mathbf{z}^{(t)}\mid\theta)$, obtain θ^{new}

replace θ^{new} with θ^{old}

end

reconstruct the feature map $\tilde{\mathbf{x}}=\mathbf{z}^{(T)}\mu^{(T)}$

end

From a perspective of time complexity, with the combined effect of convolutional layers and pooling layers, the dimension of the latent space where the latent variable $\mathbf{z}$ resides is much lower than the dimension of the sample space. Therefore, the number of Gaussian mixture models (GMM) is typically satisfied by $K\ll N$. Since the time complexity of GMM iterations is constant, the overall time complexity is $O(NK)$, which is linear with respect to the input size. This results in a significant reduction in time complexity compared to the time complexity of the Non-Local module $O\left(N^2\right)$.

From the perspective of storage space, the reconstruction process redistributes the features from high-dimensional space to a low-dimensional manifold. As a result, the reconstructed $\tilde{\mathbf{x}}$ becomes low-rank. The above process can be viewed as a decomposition of the feature map matrix $\mathbf{x}_{N\times C}$, as shown in equation (11), the storage space of the matrix can be reduced.

$$\mathbf{x}_{N\times C}=\mathbf{z}_{N\times K}\mu_{K\times C} \tag{11}$$

Finally, the residual learning and identity mapping are established using 1x1 convolutions to allow the reconstructed $\tilde{\mathbf{x}}$ to acquire the original information from the input feature map $\mathbf{x}$, thereby enhancing the feature aggregation capability.

1.2 KNN Adaptive Dual Attention

Although the GMM-based adaptive attention mechanism effectively reduces the computational complexity of self-attention and obtains a global view, enhancing the perception capability of safety helmets, the visual differences between different subcategories of safety helmets are relatively small. The feature differences mainly manifest at a fine-grained level among different safety helmets. For such fine-grained classification problems, the key lies in obtaining important feature representations.

In convolutional neural networks, each channel represents a feature map, and feature maps at different levels are responsible for extracting different types of features. In SENet, the SE attention module (SE Block) has the following form:

$$\omega=\sigma\left(f_{(\mathbf{w}_1,\mathbf{w}_2)}(g(\mathbf{x}))\right) \tag{12}$$

where $\mathbf{x}$ represents the input feature map and $\mathbf{x}\in\mathbb{R}^{C\times H\times W}$, $g(\cdot)$ denote global average pooling(GAP), the expression is given by $g(\mathbf{x})=\frac{1}{WH}\sum_{i=1,j=1}^{W,H}\mathbf{x}_{ij}$. $\sigma(\cdot)$ corresponds to the sigmoid activation function. For $f_{(\mathbf{w}_1,\mathbf{w}_2)}(\cdot)$, it can be expressed as:

$$f_{(\mathbf{w}_1,\mathbf{w}_2)}(\mathbf{y})=\mathbf{W}_2\mathrm{ReLU}\left(\mathbf{W}_1 g(\mathbf{x})\right) \tag{13}$$

where $\mathbf{W}_1\in\mathbb{R}^{C\times\frac{C}{r}}$, $\mathbf{W}_2\in\mathbb{R}^{\frac{C}{r}\times C}$. Regarding the inter-channel information interaction, the SE Block achieves local inter-channel interaction by dividing the channels into G groups. However, dividing the channels into separate groups for computation not only increases memory access costs but also loses the dependency between different groups. From equation (13), it can be observed that the dimension reduction in the non-linear activation breaks the direct correspondence between channels and channel weights. From the perspective of sparsity, fully connected layers represent dense connections, while convolutional layers represent sparse connections. Based on this, 1D convolution is introduced as a replacement for fully connected layers to capture channel attention relationships. The nearest k channels are selected as the input for 1D convolution, resulting in the creation of a novel dynamic channel attention

mechanism named KNN Channel Attention (KCA). Specifically, for the input feature map $\mathbf{x}$, it first undergoes global average pooling to obtain the aggregated feature representation $\mathbf{x}_A$, as shown below:

$$\mathbf{x}_A = GAP(\mathbf{x}) \tag{14}$$

where $\mathbf{x}_A \in \mathbb{R}^{1\times 1\times C}$. Then, 1D convolutions are used to learn the channel attention. Specifically, based on the i element in $\mathbf{x}_A$, the nearest k elements are selected as the visual feature primitive set to model the channel attention. The attention scores are then calculated using the sigmoid activation function. The calculation process is shown below:

$$\omega = \sigma\left(\mathrm{C1D}_k(\mathbf{x}_A)\right) \tag{15}$$

where $\mathrm{C1D}(\cdot)$ represents the 1D convolution, and ω denotes the attention scores. Finally, the feature map $\mathbf{x}$, is multiplied by the attention scores ω to obtain the final feature output $\tilde{\mathbf{x}}$, as shown below:

$$\tilde{\mathbf{x}} = \omega \otimes \mathbf{x} \tag{16}$$

where $\otimes$ represents the Hadamard product. Regarding the choice of the number of neighbors k, since high-dimensional/low-dimensional channels typically require longer/shorter range convolutions, the number of neighbors k is dynamically adjusted in an adaptive manner. Specifically, a band matrix $\mathbf{W}_k \in \mathbb{R}^{k\times C}$ is defined, and all channels share the parameters in $\mathbf{W}_k$, which can be expressed as follows:

$$\begin{bmatrix} w^{1,1} & \cdots & w^{1,k} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,k+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w^{C,C-k+1} & \cdots & w^{C,C} \end{bmatrix} \tag{17}$$

In the band matrix, a narrow bandwidth will weaken the channel interaction capability, while a wide bandwidth will result in inefficient computation. To address this, a non-linear mapping function $k = \varphi(C)$ is established for the channel C and the number of neighbors k, where $\varphi(\cdot)$ is defined as follows:

$$k = \left|\max\left(\log\left(C + 2^{\alpha} + \max(0,\beta)\right), \frac{C}{\gamma}\right)\right|_{odd} \tag{18}$$

where α, β, γ represents the channel adjustment factor and $|\bullet|_{odd}$ is the nearest odd number to the result. When C is small, $\mathbf{W}_k$ is denser, and modeling attention information for a small subset of channels is sufficient to capture rich attention. When C is large, the original features are greatly refined due to the exponential growth of channels, requiring more channels to participate in the interaction calculation.

Upon revisiting equation (16), $\tilde{\mathbf{x}}$ can be seen as reconstructing $\mathbf{x}$ under the influence of ω, while in equation (9), it can be seen as reconstructing $\mathbf{x}$ under the influence of $\sum_k \pi_k \gamma(\mathbf{z}_{nk})^{new} \mu_k^{new}$. Therefore, the KNN channel attention is embedded into the joint reconstruction $\mathbf{x}$ in M steps, as shown in Figure 1. For equation (9), it can be rewritten as:

$$\begin{cases} \tilde{\mathbf{x}} = \sum_{k=1}^{K} \sigma\left(C1D\left(GAP\left(\gamma(\mathbf{z}_{nk})^{new}\right)\right)\right) \otimes \mathbf{x} \\ \mathbf{x}_n^{new} = \sum_{k=1}^{K} \tilde{\mathbf{x}} \mu_k^{new} \end{cases} \tag{19}$$

From equations (18) and (19), it can be observed that by embedding the KNN channel attention, the increased parameter quantity is only at a constant level. This allows for obtaining global spatial feature information, local spatial information, and fine-grained semantic information simultaneously in linear time complexity. Finally, the constructed lightweight attention module is named KNN Adaptive Dual Attention Module (KADA Block), and its structure is illustrated in Figure 2.

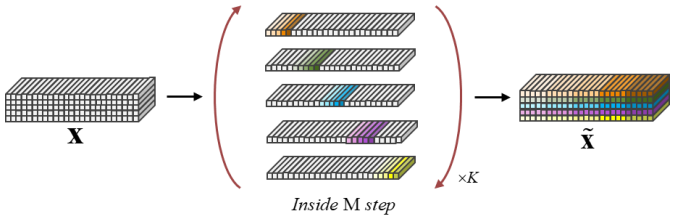


Figure 1: In the M steps, the KNN channel attention and self-attention are computed simultaneously to jointly reconstruct the input feature map.

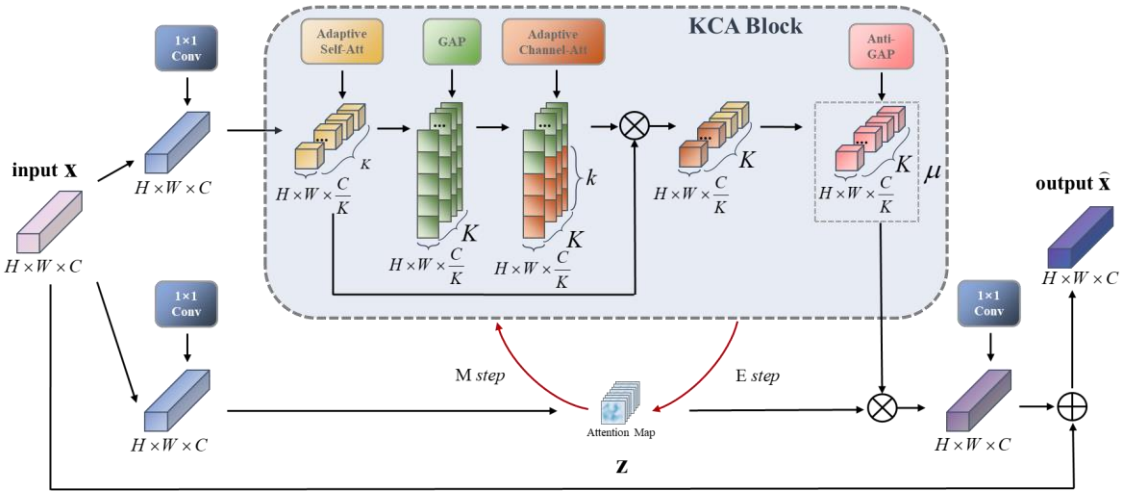


Figure 2: The principle of KNN Adaptive Dual Attention(KADA) algorithm.

2 EXPERIMENTS

2.1 Experimental Dateset

This study conducted experiments on the Safety Helmet Wearing Dataset (SHWD) [17] and the Complex Real-world Construction Site (CRCS) dataset. The SHWD dataset consists of 7581 images with 2 categories: "head" and "helmet," containing 88956 and 6817 instances, respectively. The CRCS dataset comprises 12836 images with 4 categories: "head," "helmet," "special_helmet," and "safe_helmet," containing 3285, 17871, 4008, and 3849 instances, respectively.

2.2 Implementation Details

This experiment was conducted on an NVIDIA TITANXP GPU (12GB) using PyTorch 2.2.0+cu121 as the algorithm framework, with a memory size of 128GB. During training, the batch size was set to 64, and the image size was set to 640. The optimizer used was SGD. For the SHWD dataset, due to fast convergence of the algorithm, the total number of epochs was set to 52. For the CRCS dataset, as the algorithm converged slower, the total number of epochs was set to 100. The performance evaluation metrics chosen for algorithm evaluation were Precision, Recall, Mean Average Precision at IoU threshold 0.5 (mAP 0.5), and Mean Average Precision across IoU thresholds from 0.5 to 0.95 (mAP 0.5:0.95).

2.3 Experimental Analysis

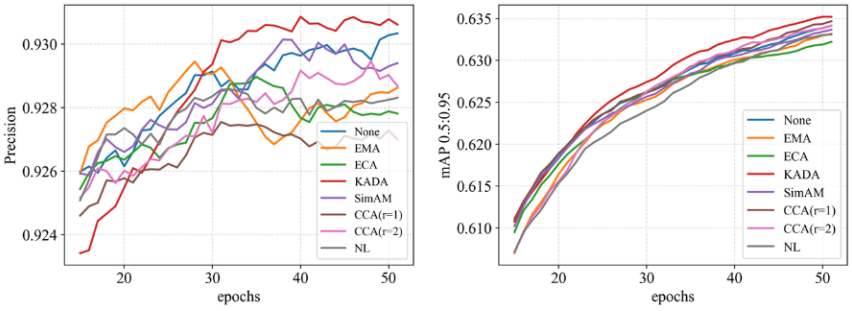
2.3.1Comparative experiments of attention mechanisms.

Figure 3 illustrates the comparison of Precision and mAP 0.5:0.95 among various attention mechanisms on the SHWD and CRCS datasets. To reduce the oscillation in the evaluation metric curves, smoothing was applied. The "None" category represents the absence of additional attention mechanisms, while the other categories represent mainstream attention modules. In the experiments, efforts were made to maintain similar computational complexity by adjusting parameters in different attention mechanisms. From Figure 3 (a) and Figure 3 (b), it can be observed that models with the addition of the KADA module outperform models with other attention mechanisms in terms of detection performance.

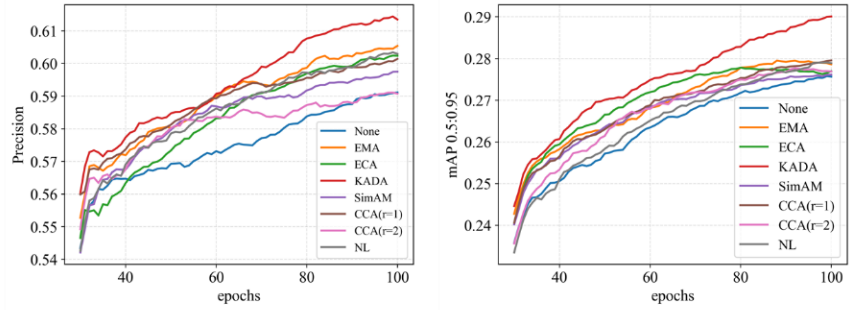
Figure 3 (a) indicates that adding other attention methods to the network on the SHWD dataset sometimes does not lead to improved performance and may even harm the model's detection capability. For example, the inclusion of EMA self-attention results in fluctuating accuracy, while the ECA channel attention method significantly damages the model's performance. Since the SHWD dataset is primarily collected from the internet, most of the samples have simple backgrounds. For these simple samples, attention mechanisms may cause the model to focus on irrelevant features. The KADA method can adaptively select and interact with important features relevant to the classes, overcoming the attention bias issue present in other attention methods.

Figure 3 (b) demonstrates that models with added attention modules are more effective than models without any attention mechanism on the CRCS dataset. The model utilizing the KADA method is more effective than models using EMA or ECA methods alone. Since the CRCS dataset consists of samples collected from the complex real world,

where workers wear different types of safety helmets, there is more background noise interference. Therefore, recognizing different types of safety helmets requires the model to have a more refined representation capability. Both the ECA method and CCA (r=1) method achieve good recognition results. The former focuses on channel interactions, while the latter focuses on long-range dependencies. However, the CCA (r=2) method restricts the improvement in accuracy. As the CCA method can be seen as computing attention maps using two independent modules in a serial manner, it suggests that this attention modeling approach is not suitable in certain scenarios. The KADA method simultaneously models self-attention and channel attention, and through an adaptive manner, couples them. This further emphasizes that capturing the indirect relationship between different modes of attention mechanisms can enhance the model's representation capability.



(a) SHWD



(b) CRCS

Figure 3: The comparison of Precision and mAP 0.5:0.95 among different attention mechanisms on the two datasets.

Table 1 presents the performance comparison of different algorithm models, where the models column highlights the proposed method in this chapter, and the metrics column highlights the best results. From the table, it can be observed that the proposed KADA method does not significantly increase the computational complexity. However, it outperforms other attention methods on both datasets. Particularly for the CRCS dataset, it achieves a significant improvement in the more stringent mAP 0.5:0.95 metric.

Table 1: The performance comparison of different models on the SHWD and CRCS datasets

model	GFLOPs	SHWD		CRCS	
		Precision	mAP0.5:0.95	Precision	mAP0.5:0.95
YOLOv5	48.2	0.9296	0.6341	0.5911	0.2756
YOLOv5+CC(r=1)	48.8	0.9267	0.6346	0.6013	0.2785
YOLOv5+CC(r=2)	55.4	0.9284	0.6341	0.5907	0.2769
YOLOv5+NL	49.2	0.9277	0.6330	0.6029	0.2788
YOLOv5+SimAM	48.2	0.9300	0.6336	0.5974	0.2761
YOLOv5+EMA	49.3	0.9286	0.6331	0.6053	0.2785
YOLOv5+ECA	48.2	0.9278	0.6321	0.6023	0.2769
YOLOv5+KADA	49.2	**0.9304**	**0.6351**	**0.6134**	**0.2900**

[a]Note: CC(r=1) represents Criss-Cross Attention with r=1 iteration, NL represents Non-Local, SimAM represents Parameter-Free Attention Mechanism.

Figure 4 illustrates the impact of various attention mechanisms on Recall and mAP 0.5. From the Recall metric curves, it can be observed that channel attention does not have a significant positive effect on recall. As training time increases, the recall rate actually decreases. This is because the inclusion of channel attention makes the model focus more on the interaction of channel features, thus improving precision, but causing a decrease in recall. Self-attention

methods, on the other hand, enhance recall by utilizing global spatial information and incorporating contextual information to reduce the rate of missed detections. Channel attention, such as ECA and SimMA, also hampers the improvement of mAP 0.5. However, the KADA method compensates for some performance loss by simultaneously computing self-attention and channel attention.

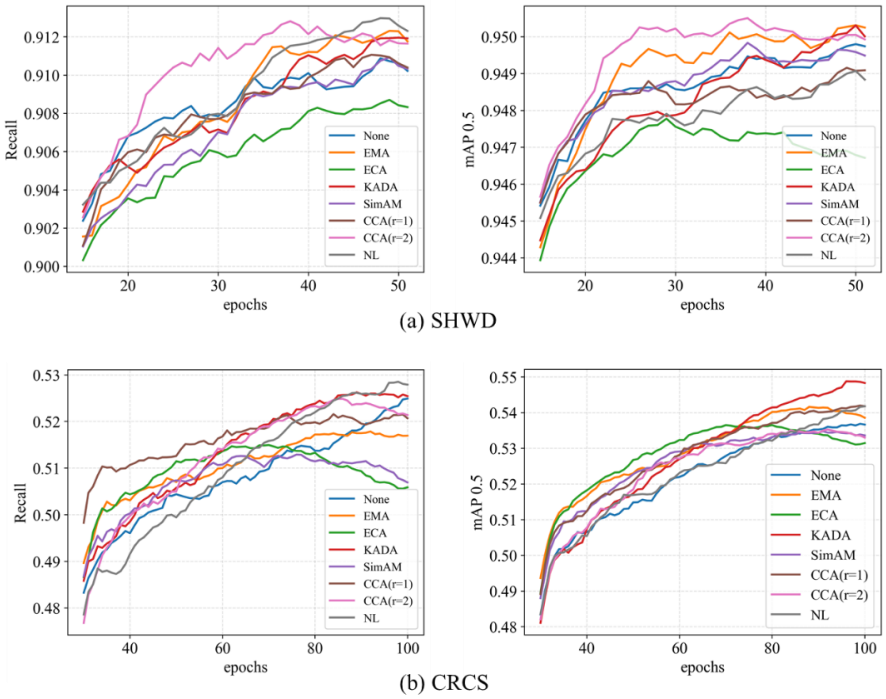


Figure 4: The impact of attention mechanisms on Recall and mAP 0.5.

2.3.2The structural analysis of the KADA method and the visualization of attention maps.

The KADA method proposed in this paper calculates self-attention and channel attention simultaneously during the computation of attention maps. In existing hybrid attention mechanisms, they are mostly designed to compute attention scores separately in a serial or parallel manner and then accumulate and fuse them [11, 18, 20, 33]. In light of this, the KADA attention is decoupled, and the effect of different positional arrangements on model performance is observed by adjusting the relative positions of the self-attention module and the channel attention module. Figure 5 depicts the computation process of attention maps for different positional arrangements.

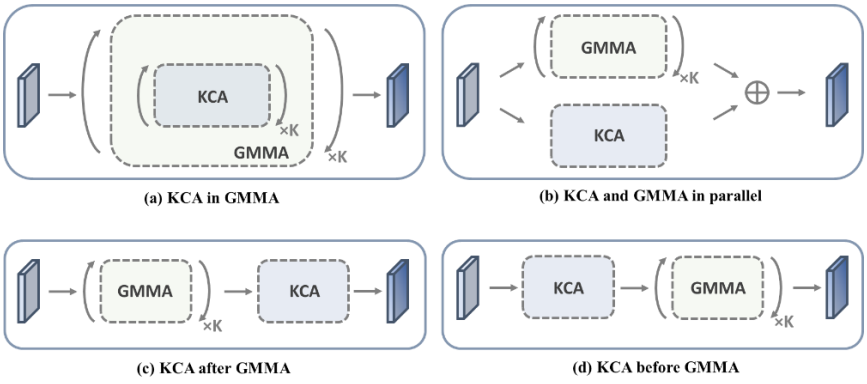


Figure 5: The positional arrangement of the KCA and GMMA attention modules.

Figure 6 presents the comparison results of Precision and mAP 0.5:0.95 for the four different arrangements of the KCA and GMMA modules. The attention embedding coupling calculation method proposed in this paper (solid red line) achieves the best results on both datasets, demonstrating that accurate attention representation can be obtained in the coupled state. For the other arrangements, on the SHWD dataset, the highest precision is obtained when KCA and GMMA are computed in parallel and fused. However, the mAP 0.5:0.95 is not as good as other arrangements. On the CRCS dataset, placing the computation of channel attention after self-attention calculation is more effective. This indicates that the order of different attention computations and their relationship with the dataset have an impact. In

some datasets, a specific computation order may yield favorable results, while in other datasets, it may not provide any gain.

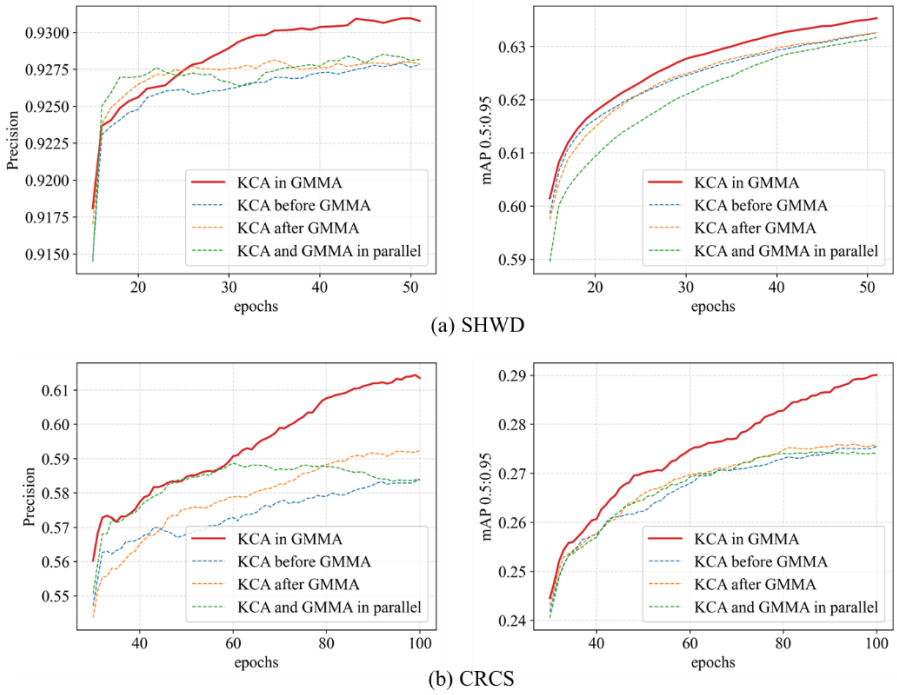


Figure 6: The performance curves of KCA and GMMA with different arrangements.

Figure 7 displays the visualization results of attention heatmaps generated by KADA, which utilizes global spatial information to reduce false detections. By incorporating self-attention, the model is able to learn semantic correlations and positional relationships between objects from the context. For instance, in the upper half of the image, without self-attention, the model tends to misclassify surveillance cameras with similar appearances as safety helmets. In the lower half, it tends to mistakenly identify the reflective parts of electric bikes as safety helmets. However, after incorporating self-attention, the semantic information of surveillance cameras/electric bikes and human heads is learned, and their positions in the feature space are far apart. As a result, the corresponding positions are hardly activated in the attention heatmap calculation, effectively avoiding false detections.

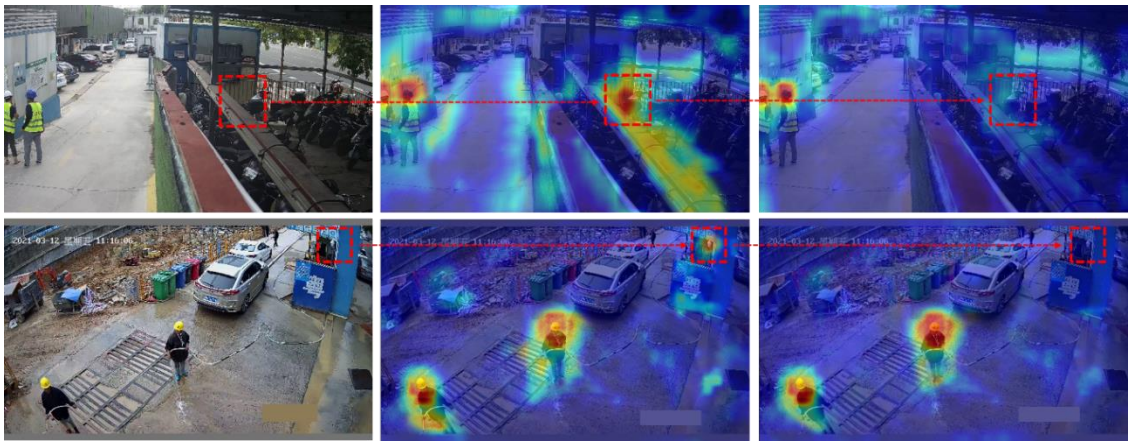


Figure 7: The application of Grad-CAM [18] allows for visual analysis to analyze the reduction of false detections through contextual dependencies in the KADA module.

Figure 8 presents the visualization results of attention heatmaps generated by KADA for recognition utilizing fine-grained channel-level feature information. The first row displays safety helmets of the "safe_helmet" category, with the most crucial distinguishing feature being the chin strap inside the helmet (as indicated by the dashed box in the figure). The second row represents the results obtained using self-attention mechanism and KADA modeling. It can be observed that when utilizing the self-attention mechanism, due to coarse feature granularity, although the safety helmet is correctly recognized, discrimination at a fine-grained level is not possible. However, when employing KADA, the crucial fine-grained semantic information used for category differentiation can be captured by the model.

Figure 8: The KADA module reduces false detections by incorporating contextual dependencies.

3 CONCLUSION

This paper proposes a lightweight dual-attention module called KADA Block. Firstly, an adaptive self-attention mechanism based on the Gaussian Mixture Model algorithm is employed to establish contextual dependencies between safety helmets and human subjects. Secondly, a KNN channel attention is utilized to capture fine-grained features of safety helmets. A KNN adaptive dual attention module is constructed, which reduces the computational complexity while obtaining both the global self-attention map and channel dependencies. Lastly, the proposed algorithm is evaluated on the SHWD dataset and CRCS dataset. Performance comparison curves of each module are plotted, and visual analysis of attention mechanisms is conducted. The experimental results demonstrate that the proposed method outperforms other state-of-the-art models discussed in the paper, highlighting the effectiveness of the proposed approach. Although KADA achieves optimal performance, the placement of the KADA module within the model is still worth discussing. While this study demonstrates that incorporating KADA along with other attention modules at the detection head yields the best results, it is still worth investigating whether KADA can be embedded in other locations, such as the backbone network.

REFERENCES

[1] Kelm A, Laußat L, Meins-Becker A, et al. Mobile Passive Radio Frequency Identification (RFID) Portal for Automated and Rapid Control of Personal Protective Equipment (PPE) on Construction Sites[J]. Automation in Construction, 2013, 36:38-52.

[2] Shrestha K, Shrestha P P, Bajracharya D, et al. Hard-Hat Detection for Construction Safety Visualization[J]. Journal of Construction Engineering, 2015, 2015(1):1-8.

[3] Fufang L, Yan C, Ming H, et al. Helmet-Wearing Tracking Detection Based on StrongSORT[J]. Sensors, 2023, 23(3):1682-1682.

[4] Farooq M, Bhutto M, Kazi A. Real-Time Safety Helmet Detection Using Yolov5 at Construction Sites[J]. Intelligent Automation & Soft Computing, 2022, 36(1):911-927.

[5] Hang S, Zhiming L, Dong R*. Partial Siamese With Multiscale Bi-Codec Networks for Remote Sensing Image Haze Removal[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61:4106516.

[6] Wang X, Ross G, Abhinav G, et al. Non-local Neural Networks[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, 18-23 June 2018:7794-7803.

[7] Xia L, Zhisheng Z, Jian W, at al. Expectation-Maximization Attention Networks for Semantic Segmentation[C]//IEEE International Conference on Computer Vision(ICCV), Seoul, October 27-Nov 2 2019:2000-2009.

[8] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, 18-23 June 2018:7132-7141.

[9] Wang Q，Wu B，Zhu P, et al. ECA-Net：efficient channel attention for deep convolutional neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 13-19 June Seattle, 2020:11531-11539.

[10] Woo S, Park J, Lee Y, et al. CBAM: convolutional block attention module[C]//European Conference on Computer Vision(ECCV), Munich, 18-14 September 2018:3-19.

[11] Fu J, Liu J, Tian H, et al. Dual Attention Network for Scene Segmentation// IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Long Beach, 15-21 Jun 2019:3146-3154.

[12] Zeiler M, Fergus R. Visualizing and Understanding Convolutional Networks[C]//European Conference on Computer Vision (ECCV), Zurich, 6-12 September 2014:818-833.

[13] Hang S, Bohui L, Zhiping D*, et al. Multi-level Feature Interaction and Efficient Non-local Information Enhanced Channel Attention for Image Dehazing[J]. Neural Networks, 2023, 163: 10-27.

[14] Zhiping D, Fang S, Sun H , et al. Outdoor Image Dehazing Based on Multi Order Channel Attention Calibration Using a Dual Discriminator Heterogeneous CycleGAN Framework[J]. ACTA ELECTONICA SINICA, 2023, 51(9): 2558-2571.

[15] Zhu J Y, Park T, Isola P, et al. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks[C]//IEEE International Conference on Computer Vision (ICCV), Venice, 22-29 October 2017:2223-2232.

[16] Hou Q, Zhou D, Feng J. Coordinate Attention for Efficient Mobile Network Design[C]//IEEE Conference on Computer Vision and pattern Recognition(CVPR), Kuala Lumpur, 19-25 June 2021:13713-13722.

[17] M. Gochoo, Safety helmet wearing dataset, in: Mendeley Data, 2021.doi:10.17632/9rcv8mm682.1.

[18] Roy A, Navab N, Wachinger C. Concurrent Spatial and Channel ‘Squeeze & Excitation’ in Fully Convolutional Networks[C]//Medical Image Computing and Computer Assisted Intervention-MICCAI, Granada, 16-20 September 2018: 421-429.