ISBN: 978-981-94-4198-3
2025 the 15th International Workshop on Computer Science and Engineering (WCSE 2025)
doi: 10.18178/wcse.2025.06.002

# Enhanced Knowledge Distillation for YOLO via Attention Mechanisms in Smart City Applications

Yanbo Wang[1] *, Shiyong Wang[1] *, Weichao LAN[1], Lili Zhang[2], Zhikun Hong[1], Hang Yao[1] , Yang Wang [1 +]

[1] Inspur Smart City Technology Co., Ltd
[2] Inspur Group Co., Ltd

**Abstract.** This paper explores the application of knowledge distillation (KD), particularly in the context of YOLO models used for real-time object detection in smart city scenarios. KD is a technique that transfers knowledge from a large teacher model to a smaller student model, enabling the latter to achieve higher accuracy while maintaining efficiency for deployment on resource-limited devices. As a popular method for distillation, Channel-wise Distillation (CWD) has been widely used for classification tasks. However, it faces limitations when applied to YOLO models for object detection, such as susceptibility to background noise. To address these issues, we introduce attention modules into the CWD framework. These modules help the model focus on relevant features, reducing background noise and improving detection accuracy, especially in complex scenes. We evaluate the performance of applying attention modules on three popular detection tasks in smart city applications, and the results demonstrate the effectiveness of our proposed method in enhancing the precision and generalization ability of YOLO models.

**Keywords:** Knowledge distillation, Object detection, YOLO, Smart city

## 1. Introduction

In deep learning, knowledge distillation (KD) is a technique used to transfer knowledge from a large and complex model, known as the teacher model, to a smaller and simpler model, referred to as the student model [1]. This process aims to capture the essential knowledge of the teacher model and embed it into the student model, which can then be deployed more efficiently on devices with limited resources, such as mobile or embedded systems.

In the context of YOLO models, which are widely used for real-time object detection in the application of smart city, knowledge distillation can be particularly beneficial. Although YOLO models are known for their speed and efficiency, some of their larger variants with better performance may still be too resource-intensive for certain applications. By applying knowledge distillation, a smaller YOLO model can inherit the knowledge from a larger, more accurate teacher model. This allows the student model to achieve higher accuracy than it would have if trained independently, while maintaining the speed and efficiency required for real-time object detection. This is crucial for applications where accurate detection is essential for ensuring both accuracy and efficiency. Moreover, knowledge distillation helps in reducing the computational load, making it feasible to deploy YOLO models on edge devices with limited resources [2-3]. This is particularly beneficial in smart city environments where real-time processing and low latency are critical. For example, in crowd management scenarios, a distilled YOLO model can quickly and accurately detect anomalies or potential safety hazards, enabling timely interventions.

Channel-wise Distillation (CWD) [4] has emerged as a widely adopted knowledge distillation method, particularly in classification tasks. It focuses on aligning the channel-wise importance of features by computing the asymmetric KL distance between the softmax distributions of the student and teacher models' activations. However, when applying CWD to YOLO models for object detection, several limitations can arise. For instance, CWD emphasis on aligning features at the channel level and is prone to background noise

* Equal Contribution.
+ Corresponding author.
*E-mail address*: wangyang19@inspur.com

in object detection tasks. This can negatively impact detection accuracy, as the model may struggle to distinguish between relevant and irrelevant features, thereby potentially leading to suboptimal performance.

To address these challenges and further enhance the precision of YOLO models, we introduce extra attention modules based on CWD. Attention mechanisms help the model focus on the most relevant parts of the input data, reducing the impact of background noise and improving the model's ability to learn critical features [5-6]. For example, by incorporating attention modules into the YOLO architecture, the model can better highlight regions with objects of interest, leading to more accurate detection. This is particularly beneficial in complex scenes where objects may be occluded or appear in cluttered backgrounds. Moreover, attention modules can optimize the feature extraction process, making it more efficient and effective. This not only improves the model's accuracy but also enhances its generalization ability to unseen data. We apply two different attention modules and evaluate the performance on three popular detection tasks in smart city application. The results demonstrate the effectiveness of the proposed method.

## 2. Related Work

KD has been widely explored in object detection to enhance the performance of lightweight models by transferring knowledge from larger, more accurate teacher models. Early work in this area includes the seminal paper by Hinton et al. [1], which first introduced the concept of distilling knowledge in neural networks. This foundational work has inspired numerous studies aiming to improve object detection models through KD.

For instance, Cao et al. [7] proposed a general distillation framework for object detectors using the pearson correlation coefficient, which has shown promising results in improving detection accuracy. Another notable approach is the focal and global knowledge distillation method by Yang et al. [8], which focuses on distilling both focal and global knowledge to enhance the performance of object detectors. For YOLO model distillation, Bharadhwaj et al. [2] utilized ensemble knowledge distillation (KD) to boost vehicle detection capabilities in YOLO Tiny. Zhao et al. [9] showcased the effectiveness of KD in lightweight models for traffic sign recognition. Meanwhile, Guan et al. [10] integrated KD with attention mechanisms in YOLOv5, resulting in enhanced performance.

These studies collectively highlight the potential of KD in improving the efficiency and accuracy of object detection models. However, challenges still remain in effectively transferring knowledge while maintaining the efficiency of the YOLO models. Thus, we propose to introduce attention mechanisms based on CMD, which help the student model focus on the most relevant features, thereby improving the precision of knowledge transfer and enhancing the overall performance.

## 3. Methodology

### 3.1. CWD Loss

CWD focuses on aligning the channel-wise distributions of feature maps between the student and teacher models. The CWD loss is calculated using the KL divergence between the normalized channel distributions of the student and teacher feature maps.

Given the feature maps S and T from the student and teacher models respectively, where $\mathrm{S} \in \mathrm{R}^{N \times C \times H \times W}$ and $\mathrm{T} \in \mathrm{R}^{N \times C \times H \times W}$, CWD loss first normalizes the feature maps along the spatial dimensions, and then applies softmax along the channel dimension to obtain the channel-wise distributions, that is,

$$\mathrm{S}_{soft\max} = \mathrm{softmax}(\frac{\mathrm{S}}{\sqrt{H \times W}}),\ \mathrm{T}_{soft\max} = \mathrm{softmax}(\frac{\mathrm{T}}{\sqrt{H \times W}}) \tag{1}$$

Finally, the CWD loss is computed as the KL divergence between the student and teacher distributions using:

$$L_{CWD} = \frac{\tau^2}{C} \sum_{i=1}^{C} \sum_{j=1}^{W \cdot H} \mathrm{T}_{soft\max}(i,j) \cdot \log(\frac{\mathrm{T}_{soft\max}(i,j)}{\mathrm{S}_{soft\max}(i,j)}), \tag{2}$$

where $\tau$ is a temperature parameter.

## 3.2. Attention Modules

To further improve the distillation performance, we propose the integration of attention modules into the CWD framework, which can significantly enhance the model's ability to focus on salient features, thereby improving the distillation process. Attention mechanisms, such as spatial attention and channel attention, have been proven to be effective in various computer vision tasks. These modules can be seamlessly integrated into the CWD framework to enrich the feature representations and improve the overall performance of the student model.

For spatial attention, the model can selectively focus on the most informative channels of the feature maps, which helps in reducing noise and emphasizing the most relevant features. In this work, we design a spatial attention module to generate a spatial attention mask that highlights important regions within the feature maps. Specifically, it contains a single convolutional layer and a sigmoid activation function is used to normalize the output of the convolutional layer. For a given feature map **X,** the max and mean values are first computed along the channel dimension to capture both the most salient and average spatial information, that is,

$$\mathrm{X}_{mean} = \frac{1}{C}\sum_{c=1}^{C}\mathrm{X}(c,:,:),\ \mathrm{X}_{\max} = \max_{c}\mathrm{X}(c,:,:) \tag{3}$$

Then, these two feature maps are concatenated along the channel dimension and passed through the convolutional layer. Finally, the output of the convolutional layer is passed through the sigmoid function to generate the spatial attention mask. This process can be formulated as,

$$\mathrm{M} = \sigma(\mathrm{Conv}(\mathrm{Concat}(\mathrm{X}_{mean}, \mathrm{X}_{\max}), \tag{4}$$

where $\sigma(\cdot)$ refers to the sigmoid function.

With respect to channel attention module, it allows the model to concentrate on specific regions of the feature maps where important information is concentrated. Specifically, the designed module consists of two fully-connected layers. First, the input feature map **X** is processed using global average pooling to reduce the spatial dimensions while retaining the channel information. Then, the output is passed through a sequence of fully connected layers to generate channel-wise attention weights. Finally, the original feature map **X** is scaled by the channel-wise attention weights using element-wise multiplication.

When integrating the designed channel and spatial attention modules into CWD loss, we first obtain the feature maps from both the student model and the teacher model, which are typically extracted from intermediate layers of the respective models. The two attention modules are then applied on the extracted feature maps for enhancement. After improving the feature maps, CWD loss is aggregated over all feature maps and channels to obtain the final loss value. This loss is then used to update the student model during training. The integration process is illustrated in Figure 1.

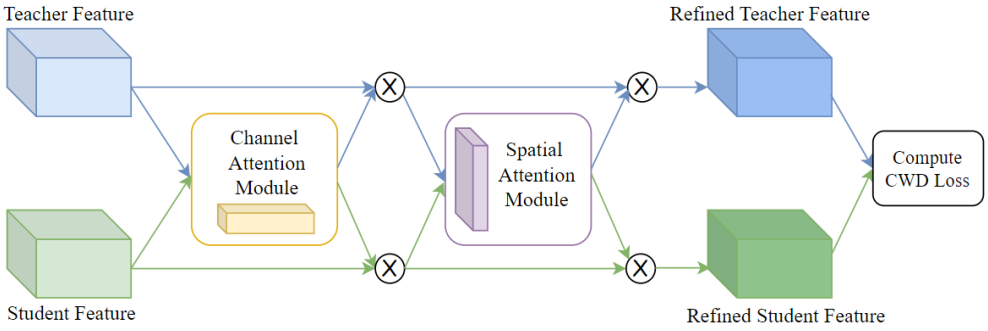


Fig. 1: Integrate the designed channel and spatial attention modules into CWD loss.

# 4. Experiments

## 4.1. Experimental Setting

**Datasets.** To comprehensively assess the efficacy of the proposed method in the context of smart city applications, we choose three representative detection scenarios, including road pothole detection, safe clothing/helmet detection and scooter violation detection. For each of these tasks, we have employed custom

datasets tailored to the specific requirements of the respective scenarios. Specifically, the road pothole dataset comprises a single class labeled as “pothole”. It consists of 3,494 images for training and 865 samples for validation. The safe clothing/helmet dataset encompasses seven distinct classes, including “self_clothes”, “safety_clothes” and “helmet”, with 9,897 images for training and 2,475 images for validation. Lastly, the scooter violation dataset includes four classes, namely “No helmet”, “One handed cycling”, “Overloadin” and “Wheeling”. It features 2,283 images for training and 134 samples for validation.

**Network and parameter settings.** When distillation, we select YOLOv8l as the teacher model and YOLOv8n as the student model for evaluation. The teacher YOLOv8l are first trained 50 epochs based on the official pretrained models, then the trained model is used to supervise the training of student YOLOv8n. For the hyparameters, we set the temperature in CWD loss as 0.5, and the weight of distillation as 1. The initial learning rate of training process is 0.001 and Adam with 0.9 momentum is applied as optimizer.

We report the detection performance, i.e., mAP@0.5 of the original teacher and student models, comparing with the performance using CWD only and CWD with the designed attention modules.

## 4.2. Results

The comparison results on three tasks are provided in Table 1. On safe clothing/helmet detection task, CWD improves the student model's performance from 0.735 to 0.740, showing that channel-wise distillation can effectively transfer knowledge from the teacher model. The CWD+Attention method further enhances the performance to 0.751, indicating that attention mechanisms help the student model focus on relevant features, thus improving detection accuracy. On scooter violation detection, the CWD+Attention method also bring significantly improvement on the performance of the student model, demonstrating the effectiveness of attention mechanisms again. With respect to road pothole, which is a relatively straightforward task compared to detecting safe clothing/helmets or scooter violations, the improvement is relatively slight. This reflects the task's inherent simplicity and the student model's ability to perform well with its own training. The effectiveness of knowledge distillation and attention mechanisms is more pronounced in tasks that are more complex and where the student model struggles to achieve high accuracy without additional guidance.

Table 1: The comparison results on three tasks.

| Teacher-YOLOv8l: 43.6M parameters, 164.8G FLOPs<br>Student-YOLOv8n: 30.0M parameters, 8.1G FLOPs | | | | | |
|---|---|---|---|---|---|
| Safe clothing/helmet | | Scooter violation | | Road pothole | |
| Method | mAP | Method | mAP | Method | mAP |
| YOLOv8l | 0.766 | YOLOv8l | 0.621 | YOLOv8l | 0.742 |
| YOLOv8n | 0.735 | YOLOv8n | 0.588 | YOLOv8n | 0.723 |
| CWD | 0.740 | CWD | 0.608 | CWD | 0.735 |
| CWD+Attention | **0.751** | CWD+Attention | **0.615** | CWD+Attention | **0.740** |


Fig. 2: Detection examples of distilled models on three tasks.

Figure 2 shows some detection models of distilled models using CWD+Attention. Overall, the results suggest that CWD+Attention generally performs better than CWD alone across all three tasks, with significant improvements observed in safe clothing/helmet and scooter violation detection. This indicates that attention mechanisms are effective for enhancing the student model's ability to focus on relevant features and improve detection accuracy. The teacher YOLOv8l consistently outperforms the other models,

confirming its superior capability. However, the student model, especially when augmented with CWD+Attention, shows promising results, demonstrating the potential for knowledge distillation to effectively transfer knowledge from complex teacher models to simpler student models for efficient deployment in smart city applications. Moreover, the student model has significantly lower computational requirements, making it more suitable for real-time processing and deployment on edge devices, which is crucial for smart city environments where low latency and real-time processing are essential.

# 5. Conclusions

This work has studied knowledge distillation techniques, specifically CWD loss and CWD enhanced with Attention mechanisms, in improving the performance of YOLO models for smart city applications. We have designed a channel attention and spatial attention module and then integrated the modules with CWD loss. The introduction of the attention mechanisms has been particularly beneficial in improving the student model's ability to focus on relevant features and reduce the impact of background noise, which is crucial for accurate object detection in complex urban environments. Our findings across three key tasks, clearly indicating that the CWD+Attention method generally outperforms the standard CWD approach. The attention modules have proven to be instrumental in enhancing the student model's precision by directing its focus towards critical regions of the input data, leading to more accurate detection. Future work could explore further enhancements to these techniques or the development of new methods specifically designed for various object detection tasks in smart city settings, potentially leading to even greater improvements in model performance and efficiency.

# 6. References


[1] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. *arXiv preprint arXiv:1503.02531*, 2015.

[2] Bharadhwaj M, Ramadurai G, Ravindran B. Detecting vehicles on the edge: Knowledge distillation to improve performance in heterogeneous road traffic. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022: 3192-3198.

[3] Setyanto A, Sasongko T B, Fikri M A, et al. Knowledge Distillation in Object Detection for Resource-Constrained Edge Computing. *IEEE Access*, 2025.

[4] Shu C, Liu Y, Gao J, et al. Channel-wise knowledge distillation for dense prediction. *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 2021, pp.5311-5320.

[5] Soydaner D. Attention mechanism in neural networks: where it comes and where it goes. *Neural Computing and Applications*, 2022, 34(16): 13371-13385.

[6] Alif M A R, Hussain M. Lightweight convolutional network with integrated attention mechanism for missing bolt detection in railways. *Metrology*, 2024, 4(2): 254-278.

[7] Cao W, Zhang Y, Gao J, et al. PKD: General distillation framework for object detectors via pearson correlation coefficient. *Advances in Neural Information Processing Systems*, 2022, 35: 15394-15406.

[8] Yang Z, Li Z, Jiang X, et al. Focal and global knowledge distillation for detectors. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022: 4643-4652.

[9] Zhao L, Wei Z, Li Y, et al. Sedg-yolov5: A lightweight traffic sign detection model based on knowledge distillation. *Electronics*, 2023, 12(2): 305.

[10] Guan P W, Zhu W X. Knowledge distillation and attention mechanism analysis of traffic sign detection. *China Automation Congress*, 2022: 2686-2691.