Transformer-Based Knowledge Distillation with Ghost Attention for Multimodal Edge-Based Smart Surveillance

Zahrah Sataar[1]*

[1]Department of Computers Techniques Engineering, College of Technical Engineering, The Islamic University, Najaf, Iraq

Abstract. In the modern era, knowledge distillation has gained attention as an important technique for edge-based smart surveillance that integrates accurate yet lightweight deployable models on resource-constrained devices. However, the existing YOLOv8 based method which integrates Coordinate Attention (CA) and Masked Generative Distillation (MGD) has faced challenges, such as relying only on infrared data, losing potential features due to excessive Learnable Dilated Convolution (LDConv) usage, and the rigidity of fixed-mask distillation. This research proposes an enhanced framework that integrates cross-architecture knowledge distillation. Infrared (IR) images are collected from the Forward Looking Infrared (FLIR) dataset, and Red Green Blue (RGB) images are collected from the Korea Advanced Institute of Science and Technology (KAIST) dataset. This followed by preprocessing using letterbox resizing, mosaic augmentation, and class-balanced sampling. In the proposed cross-architecture distillation setup, a transformer-based detector is employed as the teacher to capture long-range dependencies and contextual relations across the image, whereas a lightweight YOLOv8n optimized with Ghost Attention (GA) and a hybrid convolutional design are employed as students. Finally, Adaptive Masked Generative Distillation (A-MGD), which dynamically adjusts the mask ratio and distills multilevel features, is used to enhance knowledge transfer. The experimental results demonstrated that the proposed Transformer-teacher Knowledge Distillation for YOLOv8n student (TransKD-YOLOv8n) framework achieved higher precision (74.85%) and recall (68.90%).

1 Introduction

In recent years, Knowledge Distillation (KD) is a technology widely used in numerous applications to address privacy concerns, computational complexity, bandwidth requirements, and generalizability. Therefore, this approach provides knowledge transmission from large teacher models to smaller student models, while balancing computational costs and performance [1]. KD is a technology used to improve the performance of lightweight networks such as MobileNet and You Only Look Once (YOLO), which provide knowledge transmission by utilizing superior teacher and student models [2]. KD addresses privacy and security issues while transmitting sensitive data. Additionally, it can run in less expensive hardware systems, which require less computational power and ultimately eliminate high-performance computational systems. Furthermore, this technique has the potential for decentralized processing, which eventually enhances accurate responses, saves bandwidth, and is suitable for locations with limited availability of the Internet [3]. However, traditional KD techniques struggle in multitasking because edge surveillance frequently involves multi-task learning such as object detection, action recognition, and handling diverse data distributions. Therefore, KD is typically designed to handle a single task individually, and often struggles to adapt to changing environments with unexpected events [4]. Moreover, it faces performance issues owing to information loss, especially in high-stake applications, such as surveillance. Eventually, this ineffective knowledge transmission in the teacher-to-student model results in the loss of information gained and affects the surveillance performance [5]. To overcome these challenges, cross-modal distillation and lifelong distillation techniques have been introduced, which can easily adapt to multi-task dynamic environments. In the case of surveillance with multiple modalities, knowledge transmission occurs smoothly from the teacher model to the student model, which enhances the model's ability to handle various inputs [6]. Additionally, generative adversarial networks (GANs) have been applied to address privacy concerns and generate synthetic data without losing sensitive information. Furthermore, combining KD with optimization techniques such as quantization or pruning improves model robustness in real-world surveillance devices [7]. The state of art methods includes Bidirectional Knowledge distribution for wifi gesture recognition which is a deep learning framework which handles Non-Independent and Identically Distributed (Non-IID) data for setup in wifi gesture recognition. Moreover, this framework has the potential to adapt to a wide range of environments, such as users moving from one room to another in their home. However, this

* Corresponding author: zahrah.sataar.iu@gmail.com

framework struggles to respond quickly when dealing with smart gesture controls from users, which might lead to an increase in the power consumption [8]. Simultaneously, a pre-trained LLM (language learning model (LLM) was introduced for an accurate and powerful server environment, which was a teacher-training model. For student model training, cross-entropy loss and kullback-Leibler (KL) divergence were combined to guide the training [9]. Moreover, this model can provide fast and accurate responses when dealing with health-monitoring devices or smart assistants. However, this model might lead to unfair results that harm users owing to biased training of LLM teachers [10].

This research's key contributions are the following:

- This research introduced enhanced lightweight Transformer-teacher KD for YOLOv8n student (TransKD-YOLOv8n) framework, where a transformer-based detector serves as a teacher model and guides YOLOv8n based student model to enable the transfer of both global semantic information and local discriminative features.
- A Ghost Attention (GA) and hybrid convolution mechanism are integrated to enhance the student model, thereby ensuring the efficiency of edge deployment.
- Adaptive Masked Generative Distillation (A-MGD) is incorporated into the methodology to introduce a dynamic distillation strategy to adjust the masking ratio based on sample complexity, enabling flexible and context-aware feature reconstruction.

Further this research is organized into following sections: section 2 provides the literature works, section 3 details the entire methodology, section 4 presents the experimental results with discussions and section 5 provides the conclusion of this research.

2 Literature review

Xiancheng Cao et al. [11] presented a KD method for lightweight infrared-object detection. Initially, a forward-looking infrared (FLIR) thermal dataset was collected that contained infrared images of seven classes. Further, the dataset was converted to the You Only Look Once (YOLO) format, and classes with fewer instances were removed to improve the training process. Furthermore, a large-scale model was developed in which YOLOv8s acts as a teacher that helps in training, and YOLOv8n acts as a lightweight student model. Subsequently, Masked Generative Distillation (MGD) was used for knowledge transfer, where the student learned to reconstruct the teachers' feature maps. Later, spatial features were extracted through linear deformable convolution (LDConv) by adjusting kernel offsets and Coordinate Attention (CA) focused on important regions. These features were passed through YOLOv8n to predict the classes. Although this method is best for advanced object detection, the low resolution of thermal cameras impedes detection.

Yuanping Shi et al. [12] demonstrated a Geometry-Aligned KD (KD) for Low-Light Apple Detection. Initially, the data contained 1200 pixel-wise aligned daytime images and cycle-GAN low-light images collected from real orchard apples under lighting conditions. Furthermore, the bright images were converted to low-light images using CycleGAN to form pixel image pairs. A teacher-student framework, Knowledge Distillation with Feature Alignment (KDFA), was developed in which the teacher was trained on bright images and students learned noisy low-light images. Subsequently, knowledge is transferred through contrastive KD and feature imitation of the cross-domain alignment to preserve geometric consistency and spatial local features. Later, the student was trained on low-light images with teacher guidance through KD, and this model detected the objects. However, it improves the detection accuracy and robustness to low-light conditions but is complex when dealing with diverse data.

Nadiah AL Nomasy et al. [13] presented a KD transformer model for an intrusion-detection system. Initially, (ToN-IoT) dataset was collected from realistic IoT environments that contain normal IoT traffic and various cyberattacks. Furthermore, data were preprocessed, including cleaning, normalization, and encoding, to remove redundancy and prepare data for training. Then, a transformer-based teacher model was trained using a hybrid attention mechanism with multihead self-attention and cross-attention layers to capture temporal sequences and network features. Furthermore, a lightweight student model was trained using a selective-based KD to learn the most important features learned by the teacher model. Finally, these features are sent to the dense layer using the Softmax function to detect normal or malicious features. This method captures long-range relationships; however, because this model depends on predefined attack patterns, it is not suitable for new attacks.

Ruiping Liu Kailun Yang et al. [14] demonstrated efficient semantic segmentation using a transformer KD. Initially, the Cityscapes dataset, which contains a sequence of streets from different cities, was collected. Furthermore, these images were densely annotated to the pixel level to create a ground truth map and then preprocessed by resizing and augmenting the input images to the feed transformer model. Later, a teacher model was trained, and the student model was guided to learn by KD using both feature maps along with cross-selective fusion, which transfers spatial information and patch embeddings with embedding alignment to align sequence-level knowledge. Finally, the global-local content mixer and embedding assistant bridge the student and teacher models and improve learning. Then, the student is trained and learns from the teacher model. Finally, the student model was detected by classifying each pixel. However, this model stoplights objects that are far from the camera, and it is difficult to train the teacher model.

Geon Woo Lee et al. [15] demonstrated an enhanced automatic noise-robust speech recognition model based on the KD. Initially, the Noisy LibriSpeech corpus dataset was collected from various real-world noises

that contain clean speech with different noises. Furthermore, the data were preprocessed by cleaning and normalizing to make them efficient for training. Subsequently, an automatic speech recognition (ASR) model was trained, and acoustic tokens were created using an ASR encoder with k-means clustering, which converts frames into tokens. Furthermore, a Speech Enhancement (SE) model was trained to enhance speech when sent through the ASR encoder and tokenizer with the help of three loss functions: negative SNR loss, tokenizer loss, and encoder loss, to clean noisy speech and produce better ANR performance. Moreover, it provides better speech recognition because SE models learn from the ASR model. However, only one SE and ASR architecture limited the generalization of the results.

3 Methodology

In this section, an enhanced lightweight Transformer-teacher KD for YOLOv8n student (TransKD-YOLOv8n) framework is proposed for efficient and accurate edge-based smart surveillance. First, the Infrared (IR) images are collected from the Forward Looking Infrared (FLIR) thermal dataset, and paired Red Green Blue (RGB) with IR images are sourced from the Korea Advanced Institute of Science and Technology (KAIST) multispectral pedestrian dataset. During preprocessing, the collected images underwent letterbox resizing, mosaic augmentation, and class-balanced sampling to prevent bias and improve the generalization. At the core of the framework is a cross-architecture KD setup, where a transformer base detector acts as a teacher model to capture long-range dependencies and contextual relations, while a lightweight YOLOv8n optimized with Ghost Attention (GA) serves as a student model for edge devices. Furthermore, knowledge transfer is improved through an Adaptive Masked Generative Distillation (A-MGD) approach, which dynamically adjusts the mask ratio and distills multilevel features. An architectural diagram of the proposed methodology is shown in Figure 1.

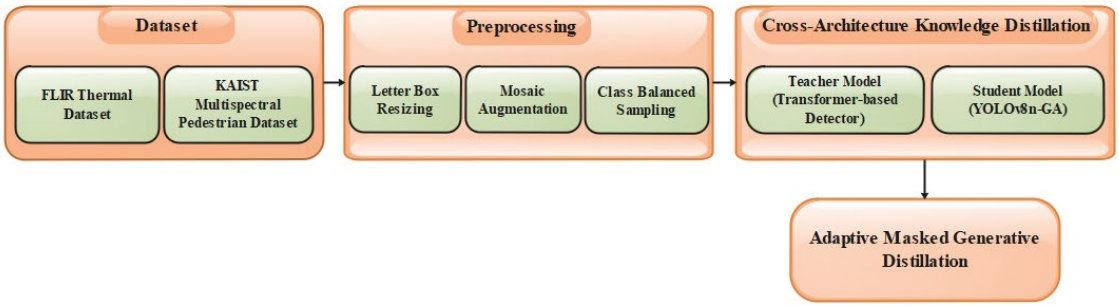

Fig. 1. Workflow of the proposed methodology.

3.1 Data collection

In this research, the input data sourced from two benchmark datasets to provide thermal and multispectral inputs for robust object detection and surveillance analysis. These datasets include These datasets include the Forward Looking Infrared (FLIR) Thermal Dataset and the Korea Advanced Institute of Science and Technology (KAIST) Multispectral Pedestrian Dataset.

3.1.1 FLIR dataset

In this research, the FLIR thermal dataset [11] is incorporated to train and validate the framework for IR object detection. It is one of the largest publicly available datasets for IR object detection and consists of high-resolution thermal images captured under various conditions such as urban traffic and environmental conditions. This dataset consists of several object categories including people, cars, bikes, motors, lights, buses, and signs. A total of 10,742 images are used for training and 1,144 images are used for validation. Initially, the original annotation was not directly compatible and the labels were converted into the YOLO annotation format. Furthermore, categories with fewer than 1000 samples were discarded from the dataset to prevent imbalances and biases.

3.1.2 KAIST multispectral pedestrian dataset

Another dataset, the KAIST multispectral pedestrian dataset [11] is incorporated into the multispectral pedestrian detection methodology. These data consist of a large collection of paired RGB-IR images, in which visible-light and thermal images are aligned and captured simultaneously. These pairs provide information, such as IR images that provide robustness under low-light conditions, whereas RGB images capture fine structural and color details. This dataset consists of diverse environments, such as city streets, campus surroundings, and residential areas, to ensure that pedestrian detection is evaluated under varying conditions. In addition, it consists of multiple pedestrian instances under daytime and night-time conditions. In this research, the KAIST dataset is incorporated to complement thermal images with aligned RGB images by enabling cross-modal feature learning.

By incorporating thermal images from the FLIR Thermal dataset and RGB-IR multispectral pedestrians, this research enhances the generalization ability. Together, these datasets are forwarded to the subsequent preprocessing stage, as explained in the next section.

3.2 Preprocessing

After data acquisition, the next stage of the methodology involved preprocessing to ensure compatibility, balance,

and enhanced generalization. This preprocessing stage included three main steps: letterbox resizing, mosaic augmentation, and class-balanced sampling.

3.2.1 Letterbox resizing

The first step in the preprocessing stage involves resizing the images collected from both the FLIR Thermal dataset and KAIST Multispectral Pedestrian dataset to meet the fixed input resolution required by the framework. This process is essential because the teaching and student models used require standardized input dimensions. In this step, letterbox resizing is employed, which preserves the original aspect ratio of each image by adding padding to ensure that the objects are not compressed or stretched. For a given input image, with (H, W) as the original dimension and the target input size (H_t, W_t). Mathematically, the scaling factor is expressed as in Equation (1):

$$s = \min\left(\frac{H_T}{H}, \frac{W_T}{W}\right) \quad (1)$$

Mathematically, the resized image dimensions are expressed in Equation (2):

$$H' = s \times H, \;\; W' = s \times W \quad (2)$$

Additionally, padding s is added to maintain the target size, as expressed by Equation (3).

$$p_h = H_T - H', p_w = W_t - W' \quad (3)$$

Where p_h and p_w denote the vertical and horizontal padding, respectively. This process ensures that the final resized image has the exact target resolution. Finally, a uniformly sized dataset consisting of images with the same resolution and preserved aspect ratios. The resized images are then forwarded to the next stage, as detailed in the next section.

3.2.2 Mosaic augmentation

The next step in preprocessing involves mosaic augmentation to improve the robustness and diversity. This technique is applied to both thermal and RGB-IR images to create a composite image by combining multiple samples, thereby forming diverse object scales, positions, and background contexts. Initially, four images are randomly selected and scaled before positioning them into a large canvas. Each image is placed in one of the four quadrants, and their bounding box annotations are adjusted accordingly. Mathematically, a mosaic image is generated using Equation (4):

$$I_m = f(I_1, I_2, I_3, I_4) \quad (4)$$

Where I_m is the mosaic image, I_i the input image, and $f(.)$ denotes spatial concatenation. The bounding box coordinates (x, y, w, h) of each object are converted using the corresponding scaling and translation matrices to align them with the new composite image. Finally, images with different scenes and objects of different categories are formed and forwarded to the next stage of class-balanced sampling.

3.2.3 Class-balanced sampling

The last step in the preprocessing is class-balanced sampling, datasets utilized in this research consist of objects from many categories, but the distribution of these categories is not standard. For instance, cars and pedestrians are seen more frequently than classes such as bicycles and signs. However, imbalance causes the model to develop bias toward leading categories while overlooking understated ones. To overcome this limitation, class-balanced sampling is used during training, here the aim is to ensure that every class contributes equally to the learning process, thereby enhancing the model's capability to detect less frequent objects. This approach works by adjusting the probability of sampling images based on the class frequency. If the frequency of class c_i in the dataset is denoted as f_i , then sampling weight w_i is presented in Equation (5).

$$w_i = \frac{1}{f_i} \quad (5)$$

The probability of selecting an image consisting of class c_i is then normalized, as expressed in Equation (6).

$$P(c_i) = \frac{1}{\sum_{j=1}^{N} w_j} \quad (6)$$

Here, N is the total number of classes, which ensures that the understated classes receive proportionally higher sampling weights, ensuring that the model learns balanced features through all categories. Therefore, a training pipeline that avoids class bias and ensures a fair illustration of all objects is achieved in this step. After all preprocessing steps, that is, letterbox resizing, mosaic augmentation, and class-balanced sampling, the dataset is now consistent, balanced, and enriched. Next, these prepared inputs are passed to the cross-architecture knowledge distillation framework, which forms the next step in the methodology.

3.3 Cross-architecture knowledge distillation

In this stage, a teacher-student knowledge distillation framework is adopted to effectively transfer knowledge from a higher-capacity model to a lightweight model. In this framework, a transformer-based detector serves as the teacher to capture long-range dependencies and global context across the image, whereas YOLOv8n enhanced with GA serves as the student.

3.3.1 Teacher model

In this section, a transformer-based detector is used as the teacher model to capture the long-range dependencies and global contextual information in an image. This transformer-based detector models the relationships between distant pixels or objects,

particularly in complex surveillance conditions where pedestrians, vehicles, and objects are present in occluded or cluttered environments. The transformer architecture is built around the multihead self-attention mechanism, where each input feature interacts with all others to calculate context-aware representations. For a given input feature sequence X, the model projects it, as expressed in Equation (7).

$$Q = XW_Q, \quad K = XW_k, \quad V = XW_v \tag{7}$$

Where, Q, K and V denote the query, key, and value matrices, respectively, W_Q, W_k and W_v denote learnable projected matrices. Mathematically, the output from self-attention is computed as expressed in Equation (8):

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

This mechanism captures both local textures and global semantics to allow the teacher to weigh the importance of each feature relative to the others. Subsequently, the outputs from multiple attention heads are concatenated and passed through the feed-forward layers to generate hierarchical feature maps. Finally, a set of feature maps and detection predictions that include both semantics and global relations are forwarded to the student model in the distillation framework, which is described in the next section.

3.3.2 Student model

In this section, YOLOv8n is used as a lightweight and efficient student model to make it suitable for deployment in edge devices. In addition, a GA used to enhance the learning capacity of the model without complexity. The GA allows the student to focus selectively on the informative parts of the input to refine the feature-extraction process. The incorporation of the GA reduces redundancy in feature maps by retaining crucial spatial details. For an intermediate feature map F, a ghost feature map F_g is generated through an inexpensive linear operation and an attention mask M is applied, as expressed in Equation (9):

$$F' = M \odot F_g \tag{9}$$

This operation ensures that the student model focuses on discriminative features while maintaining low computational cost. In addition, a hybrid convolutional design that combines standard, depthwise, and pointwise convolutions is proposed. Pointwise convolutions combine information across channels, whereas depthwise convolutions reduce Floating Point Operations Per Second (FLOPs) by operating channel-wise. Mathematically, the computational cost for the depthwise and pointwise blocks is expressed as Equation (10):

$$FLOPs = H \times W \times (C_{in}.K^2 + C_{in}.C_{out}) \tag{10}$$

Where C_{in} and C_{out} denote the input and output channels, respectively, and K denotes kernel size. This ensures that the computational overhead is reduced by maintaining the expressive capacity. Finally, an effective detector that consists of knowledge from the teacher model and is efficient for real-time inference is generated as the output. Subsequently, the student model acts as the target of the knowledge distillation process, which is described in the next section.

3.3.3 Knowledge distillation process

In this section, the knowledge distillation process, which acts as a bridge between the transformer teacher model and the lightweight YOLOv8n student model, is detailed. This process mainly focuses on enabling the student to repeat the teacher's performance, while preserving its efficiency for edge deployment. This framework applies a multilevel feature-transfer mechanism to ensure that both semantic information and structural details are effectively distilled. The distillation process consisted of two components: feature-level distillation and prediction-level distillation. In feature-level distillation, a mapping function is used to align the intermediate representations from the teacher F_T with those from student F_S. Mathematically, this mapping function can be expressed as in Equation (11):

$$L_{feat} = \|F_T - F_S\| \tag{11}$$

At the prediction level, the student is trained to replicate the output probabilities of the teacher by applying a temperature scaling parameter τ to the logit from the teacher generating soft targets, as expressed in Equations (12) and (13):

$$p_i^T = \frac{\exp\left(z_i^T/\tau\right)}{\sum_j \exp\left(z_i^T/\tau\right)} \tag{12}$$

$$L_{pred} = -\sum_i p_i^T \log\left(p_i^S\right) \tag{13}$$

Where z_i^T denotes teacher logits, and p_i^S denotes student probabilities. Finally, the total distillation loss combines both objectives with an appropriate weighting, as expressed in Equation (14).

$$L_{KD} = \alpha L_{feat} + \beta L_{pred} \tag{14}$$

Where α and β control the balance between the feature and prediction transfer. Finally, a student model that learns semantically enriched features is designed and guided by the teacher's intermediate representations and soft predictions. Next, the A-MGD, which enhances multilevel feature transfer, is discussed.

3.4 Adaptive masked generative distillation

The conventional Masked Generative Distillation (MGD) technique employs a fixed masking ratio to suppress redundant features and encourages the student

approach to rebuild missing data. Meanwhile, this method struggles with rigidity as a constant mask ratio that is unable to adapt to the variable complexity of various training samples. To address this drawback, A-MGD dynamically adjusts the masking ratio during training, allowing for a more flexible and context-aware knowledge transfer. In this method, a portion of the teacher feature map F_r is masked by utilizing an adaptive mask M_α generating a partially hidden demonstration, which is expressed in Equation (15):

$$F_T' = M_\alpha \odot F_r \tag{15}$$

Where, $\odot$ represents element-wise multiplication, the ratio of masked elements in M_α is determined by the complexity of the input, measured through entropy. Then, the student is tasked with rebuilding the full feature illustration from the partially masked teacher features. Mathematically, the reconstruction loss is given by Equation (16):

$$L_{rec} = \left\| F_s - \widehat{F_r} \right\|_2^2 \tag{16}$$

Here, F_s denotes the student feature map and $\widehat{F_r}$ denotes the reconstructed teacher feature, which enables more masking for simple samples and less masking for complex samples. Finally, the overall A-MGD loss is integrated with the features and objectives of the prediction distillation, which is described in Equation (17):

$$L_{total} = L_{KD} + \gamma L_{rec} \tag{17}$$

Here, γ signifies a weighting factor controlling the contribution of A-MGD, and the output of this process is a student approach that can learn multilevel context-aware knowledge. By influencing adaptive masking, while students inherit semantic richness and structural effectiveness, detection enhancement becomes stronger under challenging surveillance conditions. This completes the knowledge transfer. Further, the framework proceeds with the training of the model and evaluation.

4 Experimental results

This section presents the evaluation and implementation of the proposed TransKD-YOLOv8n framework. The implementation was conducted using Python programming language with the PyTorch framework, and the experiments were conducted on a system equipped with an NVIDIA A100 GPU, Intel Xeon Gold 6254, and 1TB RAM. The evaluation was performed using metrics including precision, recall, f1-score, Average Precision (AP), Mean Average Precision (mAP), Frames Per Second (FPS), and Floating-Point Operations (FLOPs). Mathematically, these metrics are expressed by Equations (18)-(23):

$$Precision = \frac{TP}{TP+FP} \tag{18}$$

$$Recall = \frac{TP}{TP+FN} \tag{19}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{20}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{21}$$

$$FLOPs = 2 \times C_{in} \times C_{out} \times K^2 \times H \times W \tag{22}$$

$$FPS = \frac{1}{T_{avg}} \tag{23}$$

Where, TP, FP and FN denote the True Positive, False Positive and False Negative, AP_i denotes the average precision for ith class; C_{in} denotes the number of input signals; C_{out} denotes the number of output channels; K^2 denotes the kernel size; $H \times W$ denotes the output feature map size; and T_{avg} denotes the average inference time.

4.1 Performance analysis

In this section, the performance of the proposed TransKD-YOLOv8n framework is evaluated by conducting a performance analysis against geometry-consistent KD and Transformer-based KD using precision, recall, and mAP. The performance analysis of the proposed TransKD-YOLOv8n framework is shown in Figure 2.

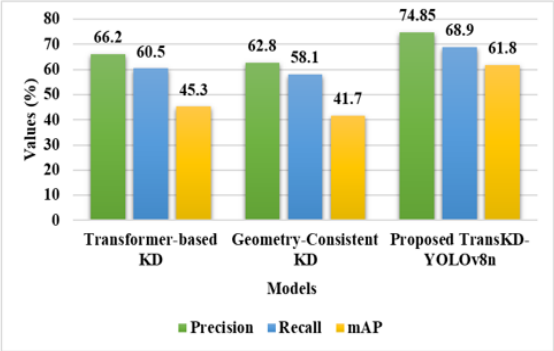

Fig. 2. Performance analysis.

Figure 2 shows the superior performance of the proposed TransKD-YOLOv8n framework in terms of precision, recall, and mAP@0.5:0.95. The bar graph clearly shows that the proposed TransKD-YOLOv8n framework outperformed the geometry-consistent KD and Transformer-based KD with higher precision (74.85%), higher recall (68.9%), and higher mAP@0.5:0.95 (61.8%) compared to the transformer-based KD and Geometry-Consistent KD based approaches.

4.2 Comparative analysis

In this section, the performance of the proposed TransKD-YOLOv8n framework is assessed by conducting a comparative analysis with the existing LKD-YOLOv8n [11] method using FLOPS, FPS, and mAP@0.5:0.95. A comparative analysis of the proposed TransKD-YOLOv8n framework is presented in Table 1.

Table 1. Comparative analysis.

Model	FLOPs (G)	FPS	mAP@0.5:0.95 (%)
LKD-YOLOv8n [11]	7.6	170	38.74
Proposed TransKD-YOLOv8n	7.9	181	61.80

From the Table 1, it is clear that the proposed TransKD-YOLOv8n framework consistently outperformed the existing LKD-YOLOv8n [11] by achieving a higher mAP@0.5:0.95 of 61.80% compared to 38.74%, higher FLOPs of 7.9 compared to 7.6 and higher FPS of 181 compared to 170. This greater performance of the proposed TransKD-YOLOv8n framework is due to the integration of cross-architecture distillation, Ghost Attention, and Adaptive Masked Generative Distillation, which enhance feature transfer and detection robustness.

4.3 Discussion

In this research, a cross-architecture knowledge distillation framework, TransKD-YOLOv8n, was proposed for edge-based smart surveillance, with the objective of obtaining accurate object detection while maintaining lightweight deployment feasibility. The existing infrared-based LKD-YOLOv8n achieved promising effectiveness, but faced significant limitations. One main limitation is that it depends solely on thermal data, which limits its adaptability to multimodal surveillance environments. Another challenge is the overuse of linear deformable convolutions that leads to the loss of fine-grained spatial features, whereas the fixed-mask distillation mechanism constrains the flexibility of knowledge transfer. These limitations limit contextual interpretation and reduce the robustness of practical edge deployment. The proposed model addresses these challenges by demonstrating a transformer-based teacher with embeddings and positional encodings, enabling the preservation of spatial relationships and long-range dependencies in the visual data. A YOLOv8n student improved with Ghost Attention, ensuring effective extraction of localized features while remaining computationally lightweight. Moreover, the AMGD mechanism dynamically adjusts the mask ratios, allowing informative regions to be distilled effectively through many levels without suppressing crucial object details. Together, these improvements yield superior results while maintaining real-time inferences. Therefore, the proposed TransKD-YOLOv8n framework introduces better reliability, robustness, and efficiency, making it more appropriate for real-world edge-based surveillance than the LKD-YOLOv8n method.

5 Conclusion

The primary objective of this research is to design an effective and reliable TransKD-YOLOv8n model for edge-based smart surveillance that obtains accurate object detection and remains lightweight for deployment. Primarily, data were collected from the FLIR Thermal dataset, which provides infrared images, and the KAIST Multispectral Pedestrian dataset, which consists of paired RGB-IR images. Next, preprocessing was performed using letterbox resizing to maintain object geometry, mosaic augmentation to improve data diversity, and class-balanced sampling to avoid bias toward dominant classes. The processed images are then fed into a cross-architecture distillation setup, where the transformer-based teacher model captured long-range dependencies and contextual features. Further, lightweight YOLOv8n students optimized with Ghost Attention and hybrid convolutions allowed computational effectiveness. To strengthen the feature transfer, an A-MGD mechanism was applied by dynamically adjusting the mask ratios to retain the informative regions. Finally, this integration enabled the proposed TransKD-YOLOv8n model to obtain higher precision (74.85%), recall (68.90%), and mAP@0.5:0.95 (61.80%) than the existing LKD-YOLOv8n approach. In future work, this model will be extended to real-time multimodal surveillance streams and optimized using lightweight transformers for deployment scalability.

References

1. D. Batic, G. Tanoni, E. Principi, L. Stankovic, V. Stankovic, S. Squartini, Interpretability and reliability-driven knowledge distillation for non-intrusive load monitoring on the edge. Expert Syst. Appl. **294**, 128837 (2025). https://doi.org/10.1016/j.eswa.2025.128837
2. A. Setyanto, T.B. Sasongko, M.A. Fikri, D. Ariatmanto, I.M.A. Agastya, R.D. Rachmanto, A. Ardana, I.K. Kim, Knowledge distillation in object detection for resource-constrained edge computing. IEEE Access **13**, 18200-18214 (2025). https://doi.org/10.1109/ACCESS.2025.3534020
3. T.M. AlZubi, H. Mukhtar, Federated knowledge distillation with 3D transformer adaptation for weakly labeled multi-organ medical image segmentation. IEEE Access **13**, 83619-83642 (2025). https://doi.org/10.1109/ACCESS.2025.3569073
4. H. Geng, D. Deng, W. Zhang, P. Ji, X. Wu, Personalized federated learning based on bidirectional knowledge distillation for WiFi gesture recognition. Electronics **12**, 5016 (2023). https://doi.org/10.3390/electronics12245016
5. T.D.T. Mai, D.Q. Vu, H.N. Phuong, T.N. Phung, EEG-SKDNet: A self-knowledge distillation model with scaled weights for emotion recognition from EEG signals. IEEE Access **13**, 140239-140250 (2025). https://doi.org/10.1109/ACCESS.2025.3594671
6. A. Santoso, Y. Surya, Maximizing decision efficiency with edge-based AI systems: advanced strategies for real-time processing, scalability, and autonomous intelligence in distributed

environments. Q. J. Emerg. Technol. Innov. **9,** 104-132 (2024)

7. D. Ollas, K. Idris, B. Thomas, Exploring adversarial attacks on TL-based edge security models, (2024).
8. N. Bakar, S.A. Azlan Ismail, F. Mokhtar, AI on edge devices for real-time image classification, (2023)
9. B. Annappa, S. Hegde, C.S. Abhijit, S. Ambesange, FedCure: A heterogeneity-aware personalized federated learning framework for intelligent healthcare applications in IoMT environments. IEEE Access **12**, 15867-15883 (2024). https://doi.org/10.1109/ACCESS.2024.3357514
10. R. Cantini, A. Orsino, D. Talia, XAI-driven knowledge distillation of large language models for efficient deployment on low-resource devices. J. Big Data **11**, 63 (2024). https://doi.org/10.1186/s40537-024-00928-3
11. X. Cao, Y. Hu, H. Zhang, LKD-YOLOv8: A lightweight KD-based method for infrared object detection. Sens. **25**, 4054 (2025). https://doi.org/10.3390/s25134054
12. Y. Shi, Y. Ma, L. Geng, L. Chu, B. Li, W. Li, KD with geometry-consistent feature alignment for robust low-light apple detection. Sens. **25**, 4871 (2025). https://doi.org/10.3390/s25154871
13. A.N. Nadiah, A. Alamri, A. Aljuhani, P. Kumar, Transformer-based KD for explainable intrusion detection system. Comput. Secur. **154**, 104417 (2025). https://doi.org/10.1016/j.cose.2024.104417
14. R. Liu, K. Yang, A. Roitberg, J. Zhang, K. Peng, H. Liu, Y. Wang, R. Stiefelhagen, TransKD: Transformer KD for efficient semantic segmentation. IEEE Trans. Intell. Transp. Syst. **25**, 20933-20949 (2024). https://doi.org/10.1109/TITS.2024.3455416
15. G.W. Lee, H.K. Kim, D.J. Kong, KD-based training of speech enhancement for noise-robust automatic speech recognition. IEEE Access **12**, 72707-72720 (2024). https://doi.org/10.1109/ACCESS.2024.3403761