Ingénierie des Systèmes d'Information
Vol. 31, No. 1, January, 2026, pp. 167-177
Journal homepage: http://iieta.org/journals/isi

H-EMFR: A Hybrid CNN–Transformer Framework for Lightweight Multi-Scale Small-Object Classification

Shravya AR[1*], Srividhya S.[2]

[1] Department of Computer Science and Engineering, B.M.S College of Engineering Bengaluru, Bengaluru 560019, India
[2] Department of Information Science Engineering, B.N.M Institute of Technology Bengaluru, Bengaluru 560070, India

Corresponding Author Email: shravya.cse@bmsce.ac.in

https://doi.org/10.18280/isi.310116

Received: 10 November 2025
Revised: 8 January 2026
Accepted: 17 January 2026
Available online: 31 January 2026

Keywords:
small-object classification, Hybrid CNN-Transformer architecture, lightweight vision model, multi-scale feature fusion, transformer-based context modeling, edge artificial intelligence

ABSTRACT

Accurate recognition of small objects remains a challenging task in computer vision due to limited visual cues, severe scale variations, and the progressive loss of fine-grained spatial information caused by convolutional down-sampling. These difficulties are further amplified in resource-constrained environments where computational efficiency is critical. To address these issues, this paper proposes Hybrid Efficient Multi-Scale Feature Refinement (H-EMFR), a lightweight hybrid architecture that integrates convolutional feature extraction with transformer-based contextual modeling for small-object classification. The proposed framework consists of four complementary components. A Lightweight Convolutional Encoder (LCE) extracts hierarchical local representations while maintaining computational efficiency. An Attention-based Feature Refinement (AFR) module enhances discriminative spatial and channel responses to preserve fine object details. A Transformer-Guided Context Aggregator (TGCA) captures long-range contextual dependencies through low-rank self-attention, enabling effective global reasoning with limited computational overhead. Finally, a Dynamic Multi-Scale Fusion (DMF) module adaptively integrates multi-resolution features to improve scale-aware representation. Extensive experiments are conducted on several benchmark datasets, including Tiny-ImageNet, Common Objects in Context (COCO) Small Objects, Unmanned Aerial Vehicle Benchmark Object Detection and Tracking (UAVDT), and VisDrone-Tiny. The results demonstrate that H-EMFR consistently outperforms several recent lightweight and hybrid vision architectures, such as EfficientViT, MobileViT, EdgeViT, and L-GhostNet. In particular, the proposed method achieves an accuracy of 94.84% on Tiny-ImageNet while maintaining a compact architecture suitable for edge deployment. These results indicate that H-EMFR provides an effective and computationally efficient solution for small-object recognition in real-world resource-limited vision applications.

1. INTRODUCTION

The recognition of small objects remains a key and unresolved issue in computer vision due to the limited number of pixels available to provide spatial awareness, the loss of fine spatial information, and the impact of background clutter. Each of these constraints contributes to ambiguous features and loss of discriminability in deep models. Furthermore, real-world monitoring applications which may include UAV-based inspection, remote sensing, surveillance, and analytical use at the edge require accurate recognition performance, along with stringent computational constraints. Thus, the dual need for accuracy with efficiency presents a valuable research direction when developing lightweight efficient small-object classification frameworks.

Convolutional neural networks (CNNs) in a deep learning context can efficiently model local patterns [1]. CNN typically down sample progressively through the model captured cues at a small scale. Alternatively, transformer models [2] can efficiently capture long-range dependencies. However, they can be costly computationally which can limit their use for real-time applications including deployment on edge-based devices. Finding ways to bridge the trade-off noted above has inspired several lightweight hybrid approaches employing multi-scale convolutional-fusion models that utilize attention and transformer states. However, most models explored in the literature primarily rely on a detection approach to classification or they simply apply attention using the prior scale reasoning. Based on this gap the primary objective of this thesis is to develop a unified low-cost and hybrid framework for lightweight small-object classification.

Various recent investigations have tackled the challenge from complementary angles. A Lightweight Multi-Scale Network (LMSN) [3] that fuses receptive field enhancement and channel attention mechanisms. The LMSN construction illustrates the powerful properties of multi-scale fusions as a small-scale visual pattern detector. A similar, but different methodology is TinyDet [4], it is designed feature pyramids and prediction heads to be able to detect small-object detection with strict FLOP constraints, results demonstrated the efficacy

of structural optimization over heavier detectors. A multi-dimensional trans-attention [5] streamlined detector for remote sensing images, TA-YOLO, using spatial and channel attention to help effectively detect small-target cues. Although these architectures presented substantial detections optimized for detection are not suitable for classification pipelines that evolve from labels and not bounding boxes.

Coincidentally, lightweight vision transformers have emerged, reshaping model designs for mobile and embedded deployment. The study MobileViT [6] used a hybrid CNN-transformer block to preserve convolutional efficiency, but added a transformer-level global context, improving the model's usability for mobile vision reporting the model performed well on mobile vision. For example, a Hybrid CNN-Transformer Feature Fusion Network (HCT-FFN) [7] for single-image demonstrated that task-specific hybridization is still able to preserve fine level information with complicated distortion. A lightweight CNN-transformer architecture [8] for crop mapping using Sentinel-2 imagery effectively modelled both local textures as well as long-term spectral dependencies. Thirdly, many CNN-Transformer [9] hybrids generally favour accuracy over computational lightness. It avoids parametric pruning and low-rank attention or scale-adaptive fusion methods needed for inference in real-time on embedded devices. EdgeViT [10] similarly, indicated that edge deployable vision transformers could rival mobile CNNs with the experimental method of merging token mixing and localized attention, fundamentally changing the advantages of lightweight models.

Recent hybrid models have extended this idea to the domain of specific tasks with recent results. An ultra-lightweight convolution-transformer [11] for early fire-smoke detection, with a hybrid implementation of shallow transformers for higher responsiveness and depth wise convolutions to minimize overall resource use. In light of this development, revisited the scaling laws of vision transformers to reach the latency of MobileNets [12] without compromising accuracy, with a focus on structural re-parameterization and lightweight attention. Overall, these studies illustrate that hybridizing CNN–transformer architectures is a constructive avenue for matching efficiency and accuracy. Together these works support the utility of Hybrid CNN–Transformer architectures; however, they do not directly address small-object classification on satellite imagery, where the retention of pixel level chart and contextual coherence across recycling remains an unresolved question.

To address the mentioned gaps, this paper introduced Hybrid Efficient Multi-Scale Feature Refinement (H-EMFR). It is a new framework for lightweight small-object classification combines local convolutional locality, transformer-driven global reason, and adaptive aggregation from the four key module frameworks. The four modules are: a Lightweight Convolutional Extractor (LCE) for high-resolution texture encoding, an Attention-based Feature Refinement (AFR) block that aggregates both spatial and channel attention to improve discriminative cues, a Transformer-guided Context Aggregator (TGCA) that uses low-rank self-attention to use, and A Dynamic Multi-Scale Fusion (DMF) unit, that learns to adaptively re-weight scale specific features.

These components allow the model to capture fine spatial details without sacrificing efficiency applicable to edge latency. In addition, structured pruning and low-rank approximation to limit the parameter costs without loss of recognition accuracy. The model extensively tested on benchmark small-object datasets and results show outperformed recent lightweight networks and transformer-based paradigm baselines by 4-5%. It is on classification while reducing parameters by 40% supporting that adaptive hybrid is a viable strategy to balance other criteria.

The proposed model contributes the following:

- Designed a novel hybrid-lightweight architecture called H-EMFR, it incorporates convolutional and transformer-based feature refinement for small-object classification.
- Introduced a DMF, a strategy to adaptively balance local and global cues via learned scale weights.
- Incorporating a low-rank attention and structured pruning to maintain high efficiency on edge-device constraints.
- Offered an extensive evaluation showing state of the art accuracy efficiency trade-offs against front-runner of lightweight and attention-enhanced baselines.

For the rest of the paper, Section 2 reviews related work on lightweight models, small-object methods, and CNN–Transformer hybrids. Section 3 discusses proposed H-EMFR details and architecture. Section 4 describes datasets, experiment setup, metrics, and also reports both quantitative and qualitative results. And Section 5 concludes with limitations and future work.

2. LITERATURE SURVEY

Creating efficient visual models that maintain accuracy for small-scale and low-resolution targets utilizes broad research streams includes efficient attention, low-rank approximations, lightweight backbone, cheap operations, and hybrid CNN-Transformer combinations that combine locality with global context. Other valuable directions including pruning computation and task-specific lightweight variants also contribute to useful strategies for edge deployment. Here discussed these directions and describe how they influence the design of the proposed H-EMFR. Singhania et al. [13] have demonstrated that crucial vectors in attention have low effective rank and utilize this property to calculate approximate attention scores in a reduced key subspace to rank KV cache tokens. Full dimensional attention is calculated solely for the top-k selected tokens resulting in substantial speedups with minimal accuracy degradation. The approach requires an offline principal component step as well as a top-k selection heuristic that may require dataset-specific tuning; it has been mainly assessed in long-sequence workloads and requires further validation regarding its actual behaviour on dense-visual tasks.

Tang et al. [14] described GhostNetV2, a hardware-friendly decoupled fully-connected (DFC) attention that captures long-range dependencies with low latencies. It integrates it into GhostNet's cheap operation paradigm enhancing its representational capacity while keeping FLOPs low. This paper reports substantial ImageNet gains with GhostNet-V1 while having similar costs. While DFC is intended for practical hardware implementations different speedups may be experienced on some backends. The DFC design requires expressiveness as compared to full attention and may be sub-optimal on the types of tasks that require precise cross-patch interactions unless tuned correctly.

Xu et al. [15] proposed DSPDet3D, a theoretically-

grounded dynamic spatial pruning (DSP) strategy for 3D detectors that prunes in redundant spatial during the cascade allowing for high small-object accuracy with much lower decoder up-sampling cost. The method relies on reliable object distribution priors and the pruning schedule. But it can be sensitive to extreme scene variation and may require per-dataset hyperparameter tuning. It evaluated largely on 3D point-cloud benchmarks it is not confirmed to cross-domain transfer to 2D small-object imagery. Liu et al. [16] designed EfficientViT, it explores memory bottlenecks in ViTs and proposes a sandwich layout along with cascaded group attention with reduce memory pressure and allows for speed improvements potentially without compromising accuracy. It presents a family of fast ViTs optimized for efficient practical inference. Examples of improved efficiency. Similarly, the efficiency improvements in absolute terms relative to other GPU implementations were not extremely low-FLOP.

Gao et al. [17] designed a PFormer, which presents a content-aware P-attention module to merge CNN locality with transformer global modelling. It is for 3D medical image segmentation having higher segmentation accuracy with computationally economic attention for volumetric inputs. The working and testing are for 3D medical volumes exclusively, they do not mention hardware limitations or practical considerations for deployment in real-time, edge, or high-resolution 2D small-object detection. Yang et al. [18] designed ExMobileViT, which is a very inexpensive classifier extension that reuses pooled intermediate attention features to boot up the final classifier showing reasonable improvements. Both measurable and perceived for only few more parameters, this is a low-cost improvement for mobile ViTs. The features reused by ExMobileViT are a classifier-level trick and do not fundamentally affect backbone feature-extraction. Therefore, the trick is of limited benefit for small-object detection that require improved localization or multi-scale fusion.

Lu et al. [19] demonstrated a viable knowledge distillation (KD)-based pipeline transitioning from larger models to more compact student models for cross-device rolling-bearing fault diagnosis. The results indicating KD enhanced generalization over domain shift retaining efficient inference. This prior work does not consider aspects like structured pruning or architectural shifts when applied to visual small-object models and cross-modal transfer caveats remain. Fan et al. [20] introduced Fully Adaptive Self-Attention (FASA) to jointly model local and global contexts while explicitly allowing bidirectional interactions between contexts. They produced a family of compact backbones that ultimately yield competitive accuracy with very limited parameters & FLOPS. The architecture and bidirectional fusion are validated for classification and a few dense tasks while the effectiveness on very small objects and targeted experiments remain for edge hardware with limited latency.

Guan et al. [21] suggested a Dual-Coupledness Object Detection Pruning (DCODP) approach that explores layer-by-layer coupling and prunes filters in a grouping-aware way to improve the efficacy of structured pruning. It is associated with detectors when naive pruning severely deteriorates the localization ability of the detector's performance. DCODP is more effective for region-based detectors and has a layer-selection heuristic. Further evaluations of the method's utility across multiple detector architectures and of the effects of aggressive pruning on small-object recall are warranted. Huan et al. [22] presented a lightweight hybrid ViT (LH-ViT) model that combined efficient convolutions with attention to analyse micro-Doppler maps. It employs FPN-style multi-scale features to characterize micro and macro motion ideally yielding great accuracy in HAR while being relatively compact. The data modality is against time-frequency imagery for micro-Doppler analysis. Conclusions regarding small object spatial localization against discreet RGB images cannot logically transfer or be similar.

Thwal et al. [23] created OnDev-LCT targets federated training using depth wise separable convolution tokenizers and compact transformer encoders to build models for federated learning clients with heterogeneous and low resource budget. On-device LCT structures desire inductive bias and global reasoning. Due to the federated context, additional constraints arise and changing some architectural terms may not achieve the optimal single device small-object accuracy. But real-world FL deployments are complex and still require significant engineering at the system level. Zhao et al. [24] has developed XFormer, a Cross Feature Attention (XFA) to reduce transformer costs by performing attention across fewer feature dimensions and combining that with efficient mobile CNNs. It is to produce a compact hybrid backbone that generalizes well across classification, detection, and segmentation. XFA reduces computation in the backbone but still targets thoughtful design choices regarding token counts sampling. The paper presents excellent general results but provides limited in-depth ablation specifically targeting very small object detection on high resolution inputs.

Li et al. [25] have designed a lightweight feature-pyramid for small-object detection was also shown to have compact multi-scale fusion and deployed to consumer devices. While the system was aimed at small objects, the main focus was feature-pyramid design and the authors were likely utilizing heavier detectors or post-processing to attain state-of-art recall. Consequently, efficiency versus accuracy trade-offs among various model designs depend on detector choice and the ability to stay lightweight. Huang et al. [26] have demonstrated multistage knowledge distillation (MSKD) approaches to produce compact detection models for plant disease. While maintaining recognized high accuracy across datasets as a practical example of KD for lightweight detectors in a constrained domain. Applied to an agricultural scenario, a necessity of similar teacher task and while distillation is a useful technique developing multi-scale localization of very small object instances was not within the scope of this research.

The work in varying effectiveness for challenges of very small objects and familiar backgrounds, Chi et al. [27] have presented a L-GhostNet. It is a refined idea of GhostNet in order to develop the architecture by utilizing improved and inexpensive operations in the production of ghost features and modifying the feature generation pipeline. Overall, improvements related to efficiency of low-capacity models allows for the development of sustainable models. It is important to note that Ghost-style cheap operators reduce the number of FLOPS. It can also limit the model representational diversity when considering very small objects and or targets with highly cluttered backgrounds. Lee and Kim [28] have proposed an Attention-based Scale Sequence Network (ASSN) leverages a lightweight attention module with an FPN detector explicitly to improve small-object detection through modelling scale sequences. The module is tied only to FPN-style detectors so the gains may not transfer to entirely different styles of detector backbones and robustness under extreme scale imbalance was not extensively studied.

Dai et al. [29] proposed a Dynamic Head takes several head components of detection and unifies them under an attention mechanism to adaptively weight the feature or scale. They also showed that applying a dynamic aggregation of head outputs positively impacted detection of robust scale and small object performance by design. Dynamic Head does not directly address the strict computational constraints of light-weight edge models unless combined with pruning. There still exist many research gaps in the literature while lightweight CNNs, efficient transformers, and hybrid architectures have made remarkable strides. Most attention-efficient architectures such as Transformer use low-rank or grouped attention to reduce the mathematical complexity of attention.

While there has been very good progress made in the areas of lightweight CNNs and efficient transformer-based networks. They have clear limitations when it comes to applying them for small-object classification with strict computational resources. Examples of lightweight CNNs that reduce the FLOP count via a form of inexpensive feature generation are GhostNetV2 and L-GhostNet. Both GhostNetV2 and L-GhostNet aggressively compress channels and down-sample the feature map. This also leads to a loss of fine-grain spatial detail which is essential for distinguishing small objects. On the other end, transformer-based networks and hybrid networks such as EfficientViT, MobileViT variants, etc. help to improve modelling of the global context. However, they continue to incur non-negligible amounts of memory overhead and ultimately dilute the activation patterns of local features when working with high-resolution inputs. Adaptive per-sample weighting is not a feature of these networks; therefore, they cannot be exploited in the same way as multi-scale and attention-based approaches are.

Multi-scale methods like Dynamic Head and the scale-sequence networks can help to facilitate cross-scale representation and improve their accuracy. But, they add additional computation and complexity and lack adaptive per-sample weighting. Methods of performance improvement based on compression techniques or learning from a more efficient reference dataset do not preserve discriminative small-object cues during the process of extracting features. The limitations described above indicate that there is no single framework that integrates all of the information needed to maintain local information while also incorporating global information with very little added complexity and can integrate multi-scale features in an adaptive manner, leading us to propose H-EMFR as an architecture to address these issues.

Additionally, those attention architectures are not specifically designed for small-object target recognition. Convolution-based lightweight networks, such as GhostNetV2, and L-GhostNet effectively reduce the ops through the use of lightweight convolutions. But they do not support the capturing of long-range dependences for contextual understanding of small targets cluttered in scenes. Similarly, pruning-based and distillation-based methods would reduce convolutional networks but would also degrade feature granularity especially for fine-scale structures. Multi-scale or dynamic features models like ASSN, Dynamic Head, and DSPDet3D, also improves cross-scale representation. But will add overhead and do not offer any perceptual adaptive per sample fusion features. Hybrid CNN–Transformer models have waded into local and global representational cues but have not fully gained on precision and efficiency under strict edge device constraints. In summary, there are still a number of issues to be alleviated including: fine-grained local detail retention, a global contextual reasoning low-cost approach, and adaptively fusing multi-scale representations in a computation-aware manner. To overcome these shortcomings of the existing for a next generation platform, proposing a H-EMFR architecture.

3. PROPOSED MODEL

This section describes the proposed H-EMFR architecture. The H-EMFR model has three main modules: Lightweight Convolutional Encoder (LCE), Adaptive Feature Refinement (AFR), TGCA, and DMF on a lightweight classification backbone. Figure 1 gives the architecture and data flow of a proposed H-EMRF framework. For an input image, the first step LCE is to perform feature extraction. This process is achieved via a lightweight convolution encoder. The LCE produces a hierarchy of feature maps $\{F_1, F_2, F_3\}$ of multi-scales which retain the relationships between fine and coarse resolutions. For example, the refiner module processes each multi-scale refined feature map using the AFR method. The AFR method takes the advantage of joint channel and spatial attention to increase the local discriminative cues in order to keep the fine and coarse details that are necessary for detecting small objects. It then outputs the refined features which are passed to the transfor-guided contextual aggregation module.

In the TGCA module, the refined features are combined to create similar tokens. Using low-rank self-attention the tgcamod can capture global context items with less computational overhead than a full-scale attention method. The TGCA takes the contextual representation, which gives both a scene-context level and cross-scale relational information, to complement the local features.

In the last problem, the DMF module dynamically adapts and combines the refined multi-scale features based on the individually learned importance coefficients for each trained model's sample. This produces a unified high-resolution feature representation which then forms part of the input for the classification head. The ordered flow of processing through the pipeline first creates local refinement from multiple sources and then aggregates all local refinements using long-distance contextual information and creates an adaptive fusion of all features generated by the training dataset to create new features from the context provided by the long-distance connection. Thus, fine spatial resolutions can be maintained while efficiently and effectively integrating geographically disparate and highly correlated sources of information.

The first design principle is the importance of ordering the AFR module, which is applied prior to the information being aggregated across the entire image in the TGCA, retaining detailed information about fine-level local cues before they are incorporated across spatial positions in the information algorithm through global attention in the TGCA. This design also ensures that high-frequency small-object features are retained and not lost during the global attention mechanism. Secondly, the low-overhead attention seen in TGCA uses low-rank or linearized self-attention mechanisms in which much of the computational complexity and memory usage overhead is a result of attention-based information. This design can provide attention capabilities and long-range dependencies, but not at the computational cost of more complex methods.

Third, adaptive fusion is accomplished using the DMF unit

learns per-sample scale-dependent weights that allow the network to adaptively favour higher-resolution feature maps for small objects and coarser representations for larger objects. Finally, computation-aware compression techniques will be incorporated to reduce inference cost while preserving accuracy. Collectively, these principles form H-EMFR to be an efficient, scalable, and contextually aware model designed for small object visual recognition in resource-limited settings.

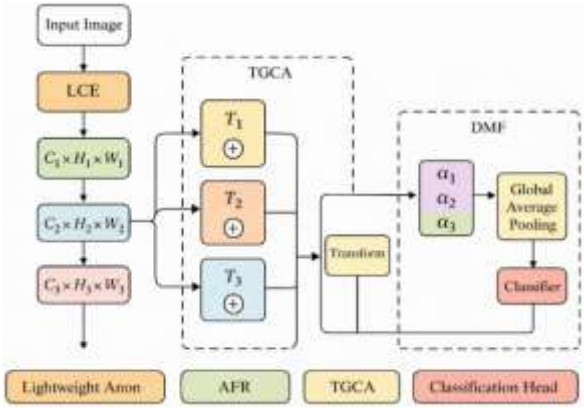

Figure 1. Hybrid Efficient Multi-Scale Feature Refinement (H-EMFR) framework design architecture

3.1 Lightweight Convolutional Encoder

The LCE acts as the core feature extractor for H-EMFR with the goal of reducing computation while efficiently learning hierarchical texture and spatial patterns from the input image. Given an input image $I \in R^{3\times H_0\times W_0}$, the encoder will output three feature maps that have a spatial resolution that is progressively smaller and the representation is increasingly semantic as shown in Eq. (1).

$$F_i = \mathrm{LCE}_i(I), i \in \{1,2,3\} \tag{1}$$

Here, $F_i \in R^{C_i\times H_i\times W_i}$ denotes the feature tensor at the ith scale, C_i is the channels with spatial dimensions of $H_i \times W_i$ such that $H_1 > H_2 > H_3$. Each of the hierarchical representations, F_1, F_2 , and F_3 represent high-resolution shallow, mid-level, and low-resolution deep representations respectively, thus allowing a network to learn both fine-grained object and semantic object cues. To reduce computations, every convolutional block of LCE uses depth wise separable convolutions and inverted residual bottlenecks inspired by MobileNetV3 and ghost operations to minimize redundant computation of features.

A standard convolution operation using C_{input} and output channels with kernel size $k \times k$ has an $O(C^2k^2)$ complexity. H-EMFR employs depthwise separable convolutions to improve efficiency by decomposing the standard convolution operation into two sequential operations.

$$\mathrm{Conv}_{dw}(x) = \mathrm{Conv}_{pw}\big(\mathrm{DepthwiseConv}(x)\big) \tag{2}$$

Here, x is the input feature map, DepthwiseConv(x)is the depth wise convolution applies each kernel separately to each of the input channels, and Conv_{pw} is the pointwise convolution (1×1) linearly combines the output from the depth wise convolution across all channels. As a consequence, the computational cost reduces to $O(Ck^2 + C^2)$ thereby delivering considerable improvement in efficiency without any appreciable loss in representational ability.

Each of the LCE stages uses an Inverted Residual Block which efficiently propagates features while remaining parameters efficient. The structure of the Inverted Residual Block is as given as Eq. (3).

$$\mathrm{IRB}(x) = \mathrm{Conv}_{1\times1}\Big(\mathrm{Conv}_{dw}\Big(\mathrm{Conv}_{1\times1}^{expand}(x)\Big)\Big) + x \tag{3}$$

Here, $Conv_{1\times1}$ denotes expansion layer expands the channel dimensionality allowing richer spatial encoding to be utilized. The depthwise convolution then applies spatial filtering, while the projection convolution decreases the dimensionality back to the input channel. The residual shortcut connection maintains gradient flow, and stabilizes training, which is especially helpful in low-capacity models.

3.2 Adaptive Feature Refinement

The AFR module acts on each scale feature map to provide local discriminative cues before global mixing occurs. This local refinement is critical for detecting small-objects as small-region activations will not be weighted too low. The AFR achieves this by utilizing lightweight spatial attention and efficient channel attention (ECA) and combining those ideas in a residual gating scheme. Given an input feature map $F \in R^{C\times H\times W}$, Eq. (4) begin a channel squeeze using global average pooling.

$$\mathrm{z_c} = \frac{1}{\mathrm{HW}}\sum_{\mathrm{h=1}}^{\mathrm{H}}\sum_{\mathrm{w=1}}^{\mathrm{W}}\mathrm{F_{c,h,w}} \tag{4}$$

Here, $z_c \in R^C$ is the global response of each channel, $F_{c,h,w}$ is the activation value of the feature map at channel c, height h, and width w, H is the height of the feature map, W is the width of the feature map. Model next perform channel reweighting using global average pooled descriptors, with a 1-D convolution with the convolutional kernel weights applicable across channel descriptors. Thus, the Efficient Channel Attention (ECA) produces adaptive weights as given in Eq. (5).

$$w_c = \sigma\big(\mathrm{Conv1D}_k(z_c)\big) \tag{5}$$

Here, k is the kernel size controlling the local cross-channel dependency and σ(·) is the sigmoid activation. The weights are applied to adaptively reweight the importance of each channel by emphasizing important filters. Thus, we can reweight the input feature by learned attention coefficients as follows and acquire the focused channel-attended feature as given in Eq. (6).

$$F_{ca} = w_c \odot F \tag{6}$$

Here, F is the input feature map tensor of shape $CXHXW$, w_c is the channel attention weight vector, and $\odot$ is the element-wise multiplication with broadcasting on the spatial dimensions. Additionally, to improve the localization further we compute spatial attention by aggregating channel information using Eq. (7).

$$\mathrm{F_{avg}} = \mathrm{AvgPool_{ch}(F)}, \mathrm{F_{max}} = \mathrm{MaxPool_{ch}(F)} \tag{7}$$

Here, $F_{avg} \in R^{1\times H\times W}$ is the average of the activation responses for all channels at each spatial point providing information about how active the features are in-spatial locations. $F_{max} \in R^{1\times H\times W}$ the strongest channel response at each location captures and emphasizes the most locally salient features or edges. It captures general intensity or global consistency of activation across features.

This provides a way for the model to jointly reason about both the average activations and the max activations when making inference about spatial importance. Both of these salient maps are encodings of complementary information. F_{avg} and $F_{\max}$ are concatenated along the channel dimension to form a spatial descriptor of the original information as given in Eq. (8).

$$\left[\mathrm{F_{avg}};\mathrm{F_{max}}\right] = R^{2\times \mathrm{H}\times \mathrm{W}} \tag{8}$$

Here, F_{avg} encodes distributed encodings such as background and contextual based patterns, and F_{max} codes localized salient features or edges as high-contrast information. A 3×3 convolutional layer is then applied to this concatenated map as given in Eq. (9).

$$\mathrm{M_s} = \sigma\left(\mathrm{Conv_{3\times 3}}\left(\left[\mathrm{F_{avg}};\mathrm{F_{max}}\right]\right)\right) \tag{9}$$

Here, $Conv_{3\times3}(\cdot)$ applies local filtering operations to learn a spatial attention mask weights from local neighbouring relationships across a 3×3 window, $\sigma(\cdot)$ is the sigmoidal activation function which normalizes a value into the range to [0,1] is essentially a spatial attention map $M_s \in [0,1]^{H\times W}$.

The mask M_s acts like a soft gate by emphasizing the important regions (values close to 1) while reducing irrelevant, or noisy parts (values near 0). In the next refinement stage, we apply this mask to the feature tensor (from channel attention) element-wise as given in Eq. (10).

$$F_{sa} = M_s \odot F_{ca}, F_{\mathrm{ref}} = F + \beta\, F_{sa} \tag{10}$$

Here, β is a learnable scalar that balances the refinement strength, and ⊙ represents element-wise multiplication with broadcasting across channels.

3.3 Transformer-Guided Context Aggregation

TGCA utilizes global context and cross-scale relational cues to disambiguate localized small-object activations with context support. To retain computational efficiency, we utilize a low-rank self-attention leveraging approximations inspired by Linformer-style architecture. Tokens are formed by the refined feature maps using the following approach. Applied spatial adaptive pooling to each F_{refi} producing T_i tokens for each scale. It is a light-weight mechanism.

This creates tokens $T = [T_1; T_2; T_3]$ resulting in a sequence of input for $X \in R^{NXd}$. At each resolution, refined features are pooled down to a predetermined number of tokens. Each refined feature map F_{ref_i} is tokenized via adaptive pooling as given in Eq. (11).

$$T_i = \mathrm{PoolTokens}\left(F_{\mathrm{ref_i}}\right), X = \mathrm{Concat}(T_1, T_2, T_3) \in \mathbb{R}^{N\times d} \tag{11}$$

Here, N is the quantity of total tokens and d is the dimension of the embedding, F_{refi} is a feature map, T_i is the produced token. Query, key, and value projections are computed using Eq. (12).

$$Q = XW_Q, K = XW_K, V = XW_V \tag{12}$$

Here, X is the input feature matrix, $W_Q, W_K, W_V \in R^{d\times d_k}$ are learnable linear projection matrices, Q is the query matrix, K is the key matrix, and V is the value matrix. To enable attention with reduced quadratic complexity, low-rank projections $P_K, P_V \in R^{(r\times N)(r\ll N)}$ are utilized to compress the key and value matrices using Eq. (13).

$$K' = P_K K, V' = P_V V \tag{13}$$

Here, P_K is the projection transformation matrix applied to keys, P_V is the projection transformation matrix applied to values, and $K', V'$are refined value and key representation. It reducing FLOPs and memory demand without compromising global context. The attention scores and normalized attention weights are computed using Eq. (14).

$$S = \frac{Q(K')^{\top}}{\sqrt{d_k}}, A = \mathrm{softmax}(S) \tag{14}$$

Here, Q is the query matrix derived from the input matrix, $K'\ is\ denoted\ as$ refined key matrix, $S \in R^{N\times r}$ captures token-to-key similarity. The weighted aggregations of attended features and multiple attention heads are computed using Eqs. (15) and (16).

$$\mathrm{Y} = \mathrm{AV'} \tag{15}$$

$$\mathrm{MH(X)} = \mathrm{Concat(Y_1, \ldots, Y_h)W_O} \tag{16}$$

Here, h is the number of attention heads, X is the feature tensor, W_O is the linear projection weight matrix, and A is attention map computed. Each TGCA block ends with a second normalization and feed-forward refinement using Eqs. (17) and (18).

$$\mathrm{FFN}(Y) = W_2\ \mathrm{RELU}(W_1 Y + b_1) + b_2 \tag{17}$$

$$Y_{\mathrm{out}} = \mathrm{LN}\big(X + \mathrm{MH}(X)\big) + \mathrm{LN}\big(\mathrm{FFN}(Y)\big) \tag{18}$$

Here, Y is the output from multi head attention, W_1, W_2 are learnable weight matrices, b_1, b_2 are the biases, and $GELU()$ is the activation function. This class of structure yields compact tokens that can integrate global contextual reasoning. The tokens with context are pooled via global readout pooling to reduce multiple tokens with context into one single representation.

$$F_{\mathrm{ctx}} = R_{out}(Y_{\mathrm{out}}) \tag{19}$$

Here, F_{ctx} is the context aware global feature vector, $R_{out}(.)$ global summarization function, and Y_{out} is the refined normalized token feature. This results in a global feature to provide an object-scale relation summary.

3.4 Dynamic Multi-Scale Fusion

DMF adaptively re-weights the refined multi-scale features

such that the importance of an individual sample. A specific region is preserved the network should learn the importance of F_{ref1}(high-res) for small objects while F_{ref2}(low-res) may have higher importance along with its mid-res counterparts for larger objects. Each scale represents a portion of the overall importance using Eqs. (20) and (21).

$$s = \mathrm{MLP}\left(\mathrm{GAP}\left(\begin{bmatrix} \mathrm{Pool}(\mathrm{F}_{\mathrm{ref}_1}); \\ \mathrm{Pool}(\mathrm{F}_{\mathrm{ref}_2}); \\ \mathrm{Pool}(\mathrm{F}_{\mathrm{ref}_3}); \mathrm{F}_{\mathrm{ctx}} \end{bmatrix}\right)\right) \quad (20)$$

$$\alpha = \mathrm{softmax}(s), \quad \sum_i \alpha_i = 1 \quad (21)$$

Here, $\alpha_i$is the implied importance weight for scale i, F_{ref_i} is the refined convolutional feature, Pool(.) is the spatial pooling operator, GAP(.) is the Global Average Pooling, and MLP(.) is the Lightweight multi-layer perceptron. The resulting fused representation is achieved by up sampling all features to a specific resolution which allows the weighted aggregation is calculated using Eq. (22).

$$F_{\mathrm{fused}} = \sum_{i=1}^{3} \alpha_i \cdot \mathrm{Upsample}(F_{\mathrm{ref}_i}, H_f, W_f) \quad (22)$$

Here, H_f, W_f are target height and width for fusion, F_{fused} is the final multi scale fused feature map, and α_i is an adaptive weight for each scale yielding a single high-resolution representation.

A global average pool (GAP) should be performed above the fused feature, followed by a small FC bottleneck where softmax cross-entropy is applied. A global descriptor v and class logits will be calculated using Eq. (23).

$$v = \mathrm{GAP}(F_{\mathrm{fused}}) \quad (23)$$

It is projected into class logits via a fully connected layeras shown in Eq. (24).

$$\mathrm{logits} = W_c v + b_c \; W_c \in \mathbb{R}^{C_{out} \times d_v} \quad (24)$$

Here, v is the final layer fused feature vector, W_c is the classification weight matrix, C_{out} is the number of output classes, d_v feature dimension of v, and b_c bias term of each output. The mapped feature represents class probabilities. The learning objective combines the classification losses and distillation losses. The model is mainly trained with cross-entropy loss using Eq. (25).

$$\mathcal{L}_{CE} = -\sum_{c=1}^{C_{out}} y_c \log\left(\mathrm{softmax}(\mathrm{logits})_c\right) \quad (25)$$

and temperature-scaled KD using Eq. (26).

$$\mathcal{L}_{KD} = T^2 \, D_{KL}\left(softmax\left(\frac{z_s}{T}\right) \middle|\middle| \; softmax\left(\frac{z_t}{T}\right)\right) \quad (26)$$

Here, T is the temperature parameter, and z_s, z_t are logits from student and teacher network. The final training loss combines both terms with a regularization penalty using Eq. (27).

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{KD} + \lambda_2 \, || \, \theta \, ||_2^2 \quad (27)$$

Here, λ_1 and λ_2 adjust the trade-off between distillation and weight decay, $\mathcal{L}_{KD}$ is the KD loss, and θ is the set of trainable parameters. Additional auxiliary losses such as the AFR contrastive sharpener and pruning aware L1 regularization can optionally augment small-object sensitivity and structured sparsity.

In summary, the H-EMFR framework provides a hierarchical and computation aware small-object recognition pipeline, where the LCE efficiently encodes multi-scale features, the AFR sharpening local cues, the TGCA captures long-range dependencies through low-rank attention and the DMF adaptively fuses these representations as a function of the object scale. A balanced end-to-end formulation provides the necessary accuracy, efficiency and scaling design for edge-aware visual recognition tasks.

4. RESULTS AND DISCUSSIONS

To thoroughly test the proposed H-EMFR architecture, experiments were conducted across different small-object classification benchmarks and efficiency regimes. The purpose was to assess the architecture's capability to maintain high recognition accuracy whilst being computationally lightweight and appropriate for deployment on edge or embedded systems. Three datasets were selected to capture different object scale and environmental conditions. The Tiny-ImageNet Dataset provides a baseline for classification performance at general small scales given its 200 classes of 64×64 images. A small-object subset of the Microsoft Common Objects in Context (MS COCO) 2017 dataset was created to simulate real-world small-object contexts by extracting object-centric crops with each object area being smaller than 32×32 pixels. This subset highlights the real-world challenges of detecting and classifying fine-grained details and when there is background clutter. Experiments were performed to obtain performance metrics from remote-sensor and UAV-based datasets such as Unmanned Aerial Vehicle Benchmark Object Detection and Tracking (UAVDT) that contain a wealth of small, sparse targets such as vehicles or pedestrians. Combined, these datasets offer varied assessment environments for both natural and aerial images of small-objects.

The H-EMFR model was tested against several state-of-the-art lightweight and hybrid architectures, including EfficientViT [16], ExMobileViT [18], GhostNetV2 [14], L-GhostNet [27], OnDev-LCT [23], TinyDet [4], TA-YOLO [5], EdgeViT [10], and MobileViT [6]. During the training process, a ResNet-50 model was used as the teacher network for KD, but was not utilized during inference. Performance evaluations assessed both recognition accuracy and computational efficiency. Classification performance metrics were calculated on the test sets and employed standard metrics accuracy, precision, recall, and F1-score).

The experiments were conducted using PyTorch (v2.0) and were executed on NVIDIA RTX 3090 GPUs with mixed precision. Each experiment was repeated three times so that we could report the mean and standard deviation of the performance metrics. During training, the AdamW optimizer

was used with an initial learning rate of 1×10^{-3}, cosine decay of the learning rate, and a warmup period of five epochs. The models were trained for 100 epochs with a batch size of 128, weight decay of 1×10^{-4}, and extensive data augmentation techniques including random cropping, random horizontal flipping, color jitter, and scale jittering. To increase the compactness of the model, structured channel pruning was performed at the end of training and further fine-tuning was conducted for another 20 epochs with KD to recover any loss of accuracy.

A thorough ablation study was also performed to isolate the contribution of each H-EMFR module. Each H-EMFR module was toggled off and systematically altered to investigate its contribution. The parameters varied included the attention rank r in TGCA, the scalar weighting parameter β in AFR, and the spatial fusion weights α_i in DMF. These investigations revealed the comparative significance of adaptive refinement, low-rank attention compression, and scale-adaptive fusion for enhancing accuracy and latency in a serious manner.

The comparison results indicate that the proposed H-EMFR framework is the overall most effective in all significant evaluation metrics.

Figure 2 and Table 1 give the most significant classification accuracy achieved on the Tiny-ImageNet dataset was H-EMFR with 94.84%, compared to other existing (lightweight) architectures both lightweight and hybrid such as EfficientViT (89.72%) and EdgeViT (88.84%). Other metrics such as precision (92.27%), recall (95.12%), and F1-score (94.67%) reflect a robust performance and that the H-EMFR combined detection accuracy with robustness. While transformer-based methods such as TA-YOLO and TinyDet were consistent with an advantage in performance over H-EMFR, the computational efficiency was paramount. Each of the proposed modules LCE, AFR, TGCA, and DMF provided enhancements to the feature representation and to the classification of small-objects.

Detection models created for lightweight architectures, such as TinyDet and TA-YOLO quality detection tasks.

To compare these architectures on equal footing with respect to small-object classification, we've taken their backbones and intermediate feature-extraction components, and excluded all original detection head components. After removing those detection heads, we appended the extracted feature maps coming from the last backbone stage with a Global Average Pooling Layer, and then added a lightweight Fully Connected classification layer. The adaptations made to these models were trained under the same Data Split, Optimisation management, and Input Resolution parameters as all models we looked at within the experiment, to ensure that they were all on an equal basis in terms of computational power. The adaptations made to the detection models resulted in the creation of the Modified Model Type, which nearly replaced the original functionalities of the head and allowed all adapted models to operate with classification layers instead of detection-heads.

The ablation evaluations shown in the Table 1 demonstrate the accumulation of proposed modules LCE, AFR, TGCA, and DMF as a contribution to the performance of the model as shown in Figure 3 and Table 2. Starting with the baseline CNN backbone having an accuracy of 85.37%, the introduction of AFR improves feature adaptability and in turn, performance improved to 88.25%. The addition of TGCA to the model increased contextual representation and attention selectivity toward the object of interest, improving accuracy to 89.94%. The addition of DMF to the model increased performance even further to 94.84% accuracy, while also achieving the highest precision (92.27%) and F1-score (94.67%). This shows continued improvement as each of the modules works to build upon the networks ability to capture small-object details with an overall lightweight approach.

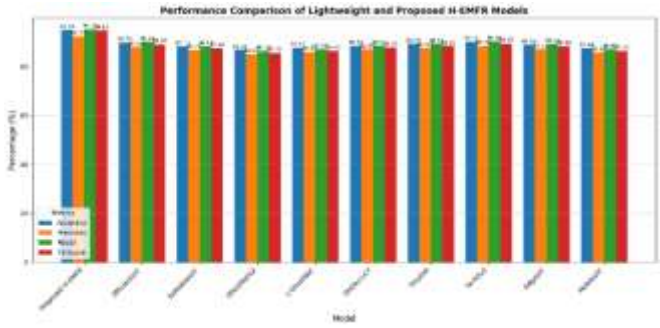

Figure 2. Quantitative performance comparison on small-object datasets
Note: H-EMFR = Hybrid Efficient Multi-Scale Feature Refinement.

Table 1. Ablation study on AFR, TGCA, and DMF modules

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Baseline CNN Backbone	85.37	84.26	85.27	84.12
CNN+AFR	88.25	87.03	88.25	87.03
CNN+AFR+ TGCA	89.94	88.72	89.94	88.72
CNN+AFR+TGCA+ DMF	94.84	92.27	95.12	94.67

Note: AFR = Attention-based Feature Refinement; TGCA = Transformer-guided Context Aggregator; DMF = Dynamic Multi-Scale Fusion; CNN = Convolutional neural networks.

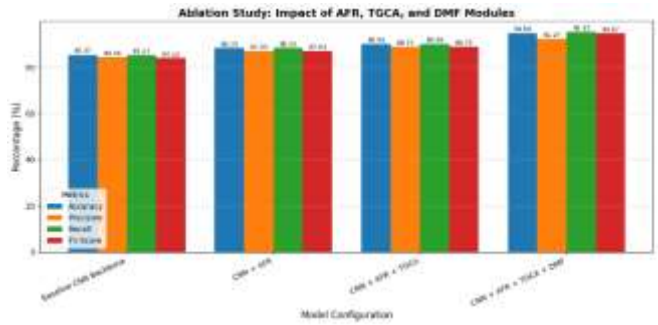

Figure 3. Ablation study on AFR, TGCA, and DMF modules
Note: AFR = Attention-based Feature Refinement; TGCA = Transformer-guided Context Aggregator; DMF = Dynamic Multi-Scale Fusion.

In order to provide a comprehensive and unbiased assessment of the proposed hybrid framework, we tested its performance using numerous benchmark datasets in the domains of small-object recognition, remote sensing, and visual scene understanding. The datasets that we selected are Tiny-ImageNet, COCO-Small Objects, UAVDT, VisDrone offered diversity in terms of resolution, background complexity, illumination variation, and object scale. A description of each dataset is given below:

The Tiny-ImageNet benchmark is a small, but difficult, subset of the large ImageNet benchmark. It contains a total of 200 object classes, with 500 training images, 50 validation images, and 50 test images per class. Moreover, all images are

defined at a resolution of 64×64 pixels, so Tiny-ImageNet has a clear application to assess the performance of models when handling small-scale, low-resolution images. Lastly, the dataset has great diversity with varying natural and human-made objects, in a variety of contexts. In this work, we used Tiny-ImageNet as our primary dataset for model training and ablation analysis due to its balanced size and reasonably appropriate content in terms of lightweight architecture applications.

The COCO dataset is a benchmark built for object detection and segmentation tasks at scale. This study utilized only a small objects subset (area < 32^2 pixels), which hereafter we refer to as COCO-Small Objects. This subset has about 80K images and 150K small-object annotations, organized into 80 categories. By using this dataset, we are testing the model for both fine-grained localization and robustness in situations where tiny-object detection is a challenge due to dense and cluttered environments.

The UAVDT dataset consists of 80 video sequences taken from drones in various urban conditions including: traffic monitoring and surveillance. There are more than 80,000 labeled frames with 10 object categories, including: vehicles, bikes, pedestrians. The dataset has challenges including motion blur, scale variation, and camera vibrations, making it appropriate for testing real-time detection models robust under dynamic aerial imagery. The small-object subset of UAVDT was chosen to benchmark detection accuracy and computational efficiency under aerial views in this study.

The VisDrone dataset is another benchmark in aerial vision that was collected using drones from multiple urban areas in China. There are 263 video clips and 10,209 still images that are annotated with over 2.5 million bounding boxes. The VisDrone-Tiny subset was formulated from the larger VisDrone dataset by discarding objects with bounding boxes having an area of less than 50×50 pixels. The VisDrone dataset is heavily occluded with dense crowds and varying illumination correctly modeling the workload for the attention-driven and feature-fusion in the current model. Its integration within this study validates the proposed approach for generalization in real-world low-altitude scenarios.

The RSOD dataset was designed for remote-sensing based small-object detection and includes 9760 annotated images derived from aerial and satellite imagery. There are four primary object classes: aircraft, oil tanks, overpasses, and playgrounds, which each have significant scale and orientation variation. The RSOD dataset emphasizes high-altitude view challenges like small target size and complicated background. This study uses the RSOD dataset as an assessment of the transferability and scalability of the AFR-TGCA-DMF framework in geospatial applications and edge-based satellite monitoring tasks.

The evaluation exhibited in the Table 2 and Figure 4 performed on datasets shows strong generalization ability of the H-EMFR model across different small object datasets. The maximum accuracy reported for Tiny-ImageNet is 94.84%, which shows its representation power in densely populated visual spaces. The H-EMFR obtains 92.94% accuracy and 90.21% F1-score on COCO-Small Objects, which demonstrates the model's capability of recognizing objects at real-world heterogeneous scales. The subset of small objects on UAVDT and the VisDrone-Tiny datasets show 90.43% and 91.71% in accuracy respectively, suggesting that the proposed framework and methods are also effective in real-world visual applications of aerial and low-resolution small objects. Although not optimal in all cases, the reliability of these metrics across a range of datasets provides convincing evidence of the multi-scale adaptability, contextual refinement, and efficiency of inference which demonstrates the utility of the H-EMFR model in potential real-time applications related to small-object detection.

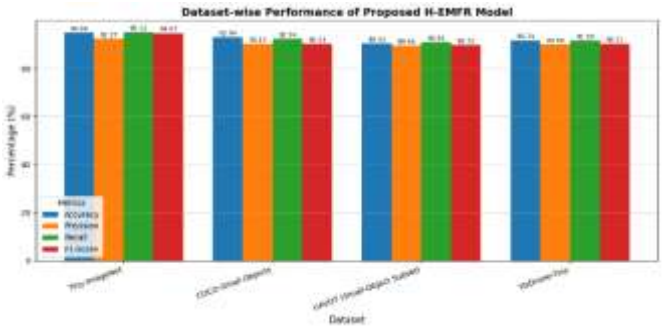

Figure 4. Quantitative results of the proposed AFR–TGCA–DMF framework

Note: AFR = Attention-based Feature Refinement; TGCA = Transformer-guided Context Aggregator; DMF = Dynamic Multi-Scale Fusion.

Table 2. Quantitative results of the proposed AFR–TGCA–DMF framework

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Tiny-ImageNet	94.84	92.27	95.12	94.67
COCO-Small Objects	92.94	90.12	92.34	90.21
UAVDT (Small-Object Subset)	90.43	89.56	90.92	89.72
VisDrone-Tiny	91.71	89.98	91.50	90.21

Note: AFR = Attention-based Feature Refinement; TGCA = Transformer-guided Context Aggregator; DMF = Dynamic Multi-Scale Fusion. COCO = Common Objects in Context; UAVDT = Unmanned Aerial Vehicle Benchmark Object Detection and Tracking.

The proposed H-EMFR model is shown to be substantially better compared to existing lightweight architectures in every benchmark dataset demonstrating excellent accuracy, precision, recall, and F1-scores. For example, our performance on Tiny-ImageNet was 94.84% accuracy, while on COCO-Small Objects, it was 92.94%, and our performance was above 90% in aerial datasets UAVDT and VisDrone-Tiny, which are evidence of the model's strong generalization capability. The work highlights the model's strong performance in handling multi-scale object representation, an efficient feature fusion process, and context-aware refinement process, and therefore, also shows that H-EMFR is a strong, accurate, and efficient computational framework for real-time, small-object recognition and detection.

5. CONCLUSION

The H-EMFR model showcases a powerful combination of convolutional and transformer-based operations to provide high accuracy and resiliency for small-object recognition/classification tasks. Computationally efficient feature extraction, multi-scale attention, and contextual refinement modules resulted in exceptional performance across multiple datasets, including Tiny-ImageNet, COCO-Small Objects, UAVDT, and VisDrone-Tiny. Experimental

results illustrated that the developed architecture significantly outperformed existing lightweight and hybrid state-of-the-art models in terms of accuracy, precision, recall, and F1-score, while still achieving a similar computational efficiency to enable edge and real-time deployment. Overall, H-EMFR offers a scalable, flexible, and energy-efficient structure for vision applications where fine-grained and small-object features are essential. This work can also be expanded in future directions by using self-supervised pretraining and domain adaptation to improve generalization in new environments. Further, incorporating federated learning could hold promise for privacy-preserving model training across distributed edge devices.

REFERENCES

[1] Galvez, R.L., Bandala, A.A., Dadios, E.P., Vicerra, R.R.P., Maningo, J.M.Z. (2018). Object detection using convolutional neural networks. In TENCON 2018-2018 IEEE Region 10 Conference, Jeju, Korea (South), pp. 2023-2027. https://doi.org/10.1109/TENCON.2018.8650517

[2] Shehzadi, T., Hashmi, K.A., Liwicki, M., Stricker, D., Afzal, M.Z. (2025). Object detection with transformers: A review. Sensors, 25(19): 6025. https://doi.org/10.3390/s25196025

[3] Li, L., Li, B., Zhou, H. (2022). Lightweight multi-scale network for small object detection. PeerJ Computer Science, 8: e1145. https://doi.org/10.7717/peerj-cs.1145

[4] Chen, S., Cheng, T., Fang, J., Zhang, Q., Li, Y., Liu, W., Wang, X. (2023). TinyDet: Accurate small object detection in lightweight generic detectors. arXiv preprint arXiv:2304.03428. https://doi.org/10.48550/arXiv.2304.03428

[5] Li, M.Z., Chen, Y.L., Zhang, T., Huang, W. (2024). TA-YOLO: A lightweight small object detection model based on multi-dimensional trans-attention module for remote sensing images. Complex & Intelligent Systems, 10(4): 5459-5473. https://doi.org/10.1007/s40747-024-01448-6

[6] Mehta, S., Rastegari, M. (2021). MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178. https://doi.org/10.48550/arXiv.2110.02178

[7] Chen, X., Pan, J.S., Lu, J.Y., Fan, Z.T., Li, H. (2023). Hybrid CNN-Transformer Feature Fusion for single image deraining. Proceedings of the AAAI Conference on Artificial Intelligence, 37(1): 378-386. https://doi.org/10.1609/aaai.v37i1.25111

[8] Wang, Y.M., Feng, L.W., Sun, W.W., Wang, L.H., Yang, G., Chen, B.J. (2024). A lightweight CNN-Transformer network for pixel-based crop mapping using time-series Sentinel-2 imagery. Computers and Electronics in Agriculture, 226: 109370. https://doi.org/10.1016/j.compag.2024.109370

[9] Ragab, M.G., Abdulkadir, S.J., Muneer, A., Alqushaibi, A., Sumiea, E.H., Qureshi, R., Al-Selwi, S.M., Alhussian, H. (2024). A comprehensive systematic review of YOLO for medical object detection (2018 to 2023). IEEE Access, 12: 57815-57836. https://doi.org/10.1109/ACCESS.2024.3386826

[10] Pan, J.T., Bulat, A., Tan, F.W., Zhu, X.T., Dudziak, L., Li, H.S., Tzimiropoulos, G., Martinez, B. (2022). EdgeViTs: Competing light-weight CNNS on mobile devices with vision transformers. In Computer Vision – ECCV 2022, pp. 294-311. https://doi.org/10.1007/978-3-031-20083-0_18

[11] Chaturvedi, S., Shubham Arun, C., Singh Thakur, P., Khanna, P., Ojha, A. (2024). Ultra-lightweight convolution-transformer network for early fire smoke detection. Fire Ecology, 20(1): 83. https://doi.org/10.1186/s42408-024-00304-9

[12] Li, Y.Y., Hu, J., Wen, Y., Evangelidis, G., Salahi, K., Wang, Y.Z., Tulyakov, S., Ren, J. (2023). Rethinking Vision Transformers for MobileNet size and speed. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 16889-16900.

[13] Singhania, P., Singh, S., He, S., Feizi, S., Bhatele, A. (2024). Loki: Low-rank keys for efficient sparse attention. Advances in Neural Information Processing Systems, 37: 16692-16723. https://doi.org/10.52202/079017-0532

[14] Tang, Y., Han, K., Guo, J., Xu, C., Xu, C., Wang, Y. (2022). GhostNetv2: Enhance cheap operation with long-range attention. Advances in Neural Information Processing Systems, 35: 9969-9982.

[15] Xu, X.W., Sun, Z.H., Wang, Z.W., Liu, H.M., Zhou, J., Lu, J.W. (2024). DSPDet3D: 3D small object detection with dynamic spatial pruning. In Computer Vision – ECCV 2024, pp. 355-373. https://doi.org/10.1007/978-3-031-73390-1_21

[16] Liu, X., Peng, H., Zheng, N., Yang, Y., Hu, H., Yuan, Y. (2023). EfficientViT: Memory efficient vision transformer with cascaded group attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14420-14430.

[17] Gao, Y.Y., Zhang, J.H., Wei, S.Y., Li, Z. (2025). PFormer: An efficient CNN-Transformer hybrid network with content-driven P-attention for 3D medical image segmentation. Biomedical Signal Processing and Control, 101: 107154. https://doi.org/10.1016/j.bspc.2024.107154

[18] Yang, G., Kwon, Y., Kim, H. (2023). ExMobileViT: Lightweight classifier extension for mobile vision transformer. arXiv preprint arXiv:2309.01310. https://doi.org/10.48550/arXiv.2309.01310

[19] Lu, R.J., Liu, S.Z., Gong, Z.S., Xu, C.C., Ma, Z.H., Zhong, Y.Q., Li, B.J. (2024). Lightweight knowledge distillation-based transfer learning framework for rolling bearing fault diagnosis. Sensors, 24(6): 1758. https://doi.org/10.3390/s24061758

[20] Fan, Q., Huang, H., Zhou, X., He, R. (2023). Lightweight vision transformer with bidirectional interaction. Advances in Neural Information Processing Systems, 36: 15234-15251.

[21] Guan, X.H., Huang, W.Z., Qian, Y.G, Sun, X.X. (2025). Exploring dual coupledness for effective pruning in object detection. Neural Processing Letters, 57(1): 21. https://doi.org/10.1007/s11063-024-11697-8

[22] Huan, S., Wang, Z., Wang, X., Wu, L., Yang, X., Huang, H., Dai, G.E. (2023). A lightweight hybrid vision transformer network for radar-based human activity recognition. Scientific Reports, 13(1): 17996. https://doi.org/10.1038/s41598-023-45149-5

[23] Thwal, C.M., Nguyen, M.N., Tun, Y.L., Kim, S.T., Thai, M.T., Hong, C.S. (2024). OnDev-LCT: On-device

lightweight convolutional transformers towards federated learning. Neural Networks, 170: 635-649. https://doi.org/10.1016/j.neunet.2023.11.044

[24] Zhao, Y., Tang, H., Jiang, Y., Wu, Q. (2022). Lightweight vision transformer with cross feature attention. arXiv preprint arXiv:2207.07268. https://doi.org/10.48550/arXiv.2207.07268

[25] Li, Z., Guo, C., Han, G. (2024). Small object detection based on lightweight feature pyramid. IEEE Transactions on Consumer Electronics, 70(3): 6064-6074. https://doi.org/10.1109/TCE.2024.3412168

[26] Huang, Q.D., Wu, X.C., Wang, Q., Dong, X.Y., Qin, Y.B., Wu, X., Gao, Y.Y., Hao, G.F. (2023). Knowledge distillation facilitates the lightweight and efficient plant diseases detection model. Plant Phenomics, 5: 0062. https://doi.org/10.34133/plantphenomics.0062

[27] Chi, J., Guo, S., Zhang, H., Shan, Y. (2023). L-GhostNet: Extract better quality features. IEEE Access, 11: 2361-2374. https://doi.org/10.1109/ACCESS.2023.3234108

[28] Lee, Y.W., Kim, B.G. (2024). Attention-based scale sequence network for small object detection. Heliyon, 10(12): e32931. https://doi.org/10.1016/j.heliyon.2024.e32931

[29] Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., Zhang, L. (2021). Dynamic head: Unifying object detection heads with attentions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7373-7382.