

Article

DASH-YOLO: A Dynamic Adaptive Sparse Hierarchical Network for Small Object Detection in Transmission Line Fittings

Yue Jiang ¹, Chaobing Zheng ² , Yuwen Li ^{3,*}, Pai Xu ³ and Wentao Huang ⁴

¹ Hubei University of Technology Detroit Green Technology Institute, Hubei University of Technology, 28 Nanli Road, Wuhan 430068, China; 2311621329@hbut.edu.cn

² School of Electronic Information, Wuhan University of Science and Technology, Wuhan 430081, China; zhengchaobing@wust.edu.cn

³ School of Electrical Engineering, Jiangxi University of Water Resources and Electric Power, 289 Tianxiang Road, Nanchang 330099, China; juwpxupai@outlook.com

⁴ School of Electrical and Electronic Engineering, Hubei University of Technology, 28 Nanli Road, Wuhan 430068, China; 20181094@hbut.edu.cn

* Correspondence: yuwen.li@juwp.edu.cn

Abstract

Power transmission line monitoring is critical for grid stability. However, existing object detection methods remain challenged by small-scale fittings in aerial transmission-line images because of feature sparsity, complex backgrounds, and computational constraints. In this paper, a YOLO-based dynamic adaptive sparse hierarchical network (DASH-YOLO) is proposed to address these issues. DASH-YOLO integrates dynamic adaptive computation with hierarchical attention to improve small-fitting feature representation while reducing redundant computation. Specifically, a Dynamic Sparse Hyper Graph (DSHG) module is introduced to regulate feature interaction density according to scene complexity, allowing relational modeling to focus on informative feature nodes. In addition, a Hierarchical Attention (HA) module adopts a coarse-to-fine refinement strategy to enhance global semantic awareness and local spatial precision. Experiments on a real-world transmission-line fitting dataset show that DASH-YOLO achieves 90.37% mAP@0.5 and 74.74% mAP@0.5:0.95, improving over YOLOv8x by 4.63 and 4.62 percentage points, respectively, while also reaching 90.80% recall, 85.49% precision, and 62.99 FPS under the same validation protocol.

Keywords: small object detection; transmission line fittings; dynamic sparse hypergraph; hierarchical attention

MSC: 68T45



Academic Editor: Dan Vilenchik

Received: 9 June 2026

Revised: 8 July 2026

Accepted: 10 July 2026

Published: 11 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The reliable operation of power transmission infrastructure is fundamental to modern society, as transmission line failures can lead to large-scale blackouts and significant economic losses. Regular inspection and maintenance of transmission line components, particularly small fittings such as various types of wire clamps, dampers and connectors, are essential for ensuring grid stability. Traditional manual inspection methods, although effective, suffer from high labor costs, safety risks, limited coverage, and inconsistent inspection quality [1]. With the development of unmanned aerial vehicles (UAVs) equipped with high-resolution cameras, transmission line inspection has been significantly enhanced

by enabling rapid and large-scale monitoring [2]. However, accurate detection of small fittings in aerial images remains a challenging task.

Small object detection in transmission line imagery has three primary challenges. First, the inherent scale disparity between small fittings and high-resolution inspection images results in limited pixel representation, inducing severe feature sparsity in deep network layers as spatial resolution progressively decreases [3]. Second, complex backgrounds typical of transmission line environments introduce substantial noise and false positive sources, including tree branches, birds, atmospheric conditions, and structural components that share visual similarities with genuine fittings [4]. Third, the limited visual information of small objects fundamentally constrains detection performance, as available features decrease quadratically while relative noise increases with diminishing object size [5].

To address these challenges, researchers have proposed various approaches. Multi-scale feature fusion techniques such as Feature Pyramid Networks (FPNs) [6] and Path Aggregation Networks (PANets) [7] are designed to preserve spatial details while integrating semantic information. Attention mechanisms including CBAM [8], ECA [9], and GAM [10] have been incorporated to enhance feature discrimination. Recent YOLO-based detectors have further improved the balance for UAV inspection scenarios. However, existing methods still face critical limitations as follows. First, uniform attention mechanisms fail to adapt to varying scene complexity, leading to computational waste in simple regions and insufficient processing in complex areas [11]. Second, traditional multi-scale fusion strategies struggle with effective information aggregation across disparate scales [12]. Third, the trade-off between detection accuracy and inference speed remains inadequately balanced for real-time applications.

To overcome these limitations, this paper proposes DASH-YOLO, a dynamic adaptive sparse hierarchical network for small fitting detection in transmission-line inspection images. The key novelty of DASH-YOLO is that it does not simply add a generic attention block to YOLOv8x; instead, it introduces a complexity-aware relational modeling strategy and a coarse-to-fine attention refinement strategy. Specifically, the Dynamic Sparse Hyper Graph (DSHG) module reshapes the deepest feature map into spatial nodes and uses a small set of learnable hyperedge prototypes to aggregate and propagate long-range feature relationships. A learned complexity score further generates a density coefficient that adaptively controls the residual contribution of hypergraph-enhanced features. In this way, the model can strengthen complex scenes containing small fittings while avoiding dense pairwise interaction over all spatial positions. Subsequently, the Hierarchical Attention (HA) module first obtains low-resolution global semantic guidance and then performs full-resolution local refinement using depth-wise spatial attention. The refined features are finally fed into a standard multi-scale detection head for bounding box regression and classification. The main contributions of this work are summarized as follows:

(1) A Dynamic Sparse Hyper Graph module (DSHG) is proposed to model higher-order and long-range dependencies among spatial feature nodes through learnable hyperedge prototypes. Compared with dense pairwise interaction, the proposed prototype-mediated propagation provides efficient relational reasoning and is further modulated by a learned density gate.

(2) A Hierarchical Attention (HA) mechanism is designed to combine global semantic guidance and local spatial refinement in a coarse-to-fine manner, thereby improving the discrimination and localization of small fittings under complex backgrounds.

(3) A high-resolution transmission-line fitting dataset with detailed annotations is constructed to support systematic evaluation of small fitting detection in real-world UAV inspection scenarios.

2. Related Works

2.1. Evolution of Power Line Inspection Methods

Power transmission line inspection has evolved dramatically over the past decade, transitioning from labor-intensive manual methods to sophisticated automated systems. Early computer vision approaches relied on handcrafted features and traditional machine learning classifiers [13], achieving limited success due to the complexity of real-world scenarios. The introduction of deep learning revolutionized transmission line inspection, with Faster R-CNN [14] and Mask R-CNN-based methods [15] achieving notable improvements in detection accuracy. Zhang et al. proposed an enhanced R-CNN framework [16] specifically designed for insulator detection, incorporating multi-scale feature fusion and contextual information. However, two-stage detectors suffer from computational overhead, making real-time deployment challenging for UAV-based inspection systems.

The YOLO family [17–22] has emerged as a preferred choice for real-time power line inspection due to its single-stage architecture and excellent balance. Recent works have adapted YOLO variants for transmission line component detection, incorporating domain-specific optimizations such as specialized anchor designs and attention mechanisms [23]. Despite these advances, detecting small fittings remains challenging, as standard architectures struggle to maintain sufficient spatial resolution while capturing semantic information [24].

More recent domain-specific studies further indicate that transmission-line inspection is not merely a generic small-object detection task. Power-line asset datasets such as STN PLAD include towers, insulators, spacers, tower plates, and Stockbridge dampers captured from high-resolution UAV images, showing large variations in scale, orientation, illumination, and background clutter [25]. Recent component-detection work has also adapted YOLO models with anchor refinement, attention modules, and focal loss to detect key transmission-line components under class imbalance and small-target conditions [26]. Compared with these studies, the present work focuses on fine-grained fittings, including clamps, dampers, and multiple connector categories, where the targets are visually similar and often occupy only a small image area. Therefore, DASH-YOLO is positioned as a fitting-oriented detector rather than a general power-line asset detector.

2.2. Small Object Detection Techniques

The challenge of detecting small objects has spawned extensive research into specialized architectures and training strategies. Feature Pyramid Networks (FPNs) pioneered top-down multi-scale fusion to enhance spatial resolution in deep layers, while Path Aggregation Network (PANet) and Bidirectional FPN (BiFPN) [27] improved cross-scale information flow via bidirectional pathways. These methods are widely adopted in YOLO-based power line inspection systems (e.g., YOLOv5/v8 neck architectures) but process all spatial regions uniformly, leading to excessive computation on uninformative backgrounds (e.g., sky regions in transmission line imagery) while underprioritizing critical fitting areas [28].

Recent advances in attention mechanisms have shown particular efficacy for small object detection. The integration of Efficient Multi-Scale Attention (EMA) [29] modules and Spatial Awareness Modules (SAM) [30] has enhanced feature discrimination for small targets. Dynamic convolution approaches adapt kernel parameters [31] based on input features, providing content-aware processing. However, these methods typically apply attention uniformly across all spatial locations, failing to account for varying scene complexity and leading to computational inefficiency.

Multi-scale hybrid attention approaches have shown particular promise, effectively utilizing similarity relationships between objects of various sizes to guide small object

detection [32]. The hierarchical application of attention at different feature scales enables progressive refinement, though existing implementations often lack explicit mechanisms for handling the extreme scale variations characteristic of transmission line imagery [33].

2.3. Attention Mechanisms in Object Detection

Attention mechanisms have emerged as powerful tools for enhancing feature discrimination in object detection tasks. Channel attention modules like Squeeze-and-Excitation (SE) [34] blocks recalibrate channel-wise feature responses, emphasizing informative channels while suppressing less relevant ones. Spatial attention mechanisms including Convolutional Block Attention Module (CBAM) [35] and Global Context Network (GCNet) [36] model spatial dependencies to highlight salient regions. Recent works have explored coordinate attention [37] and efficient channel attention [38], achieving improved performance with reduced computational cost. Hybrid attention mechanisms that combine channel and spatial attention have become prevalent for their comprehensive feature enhancement capabilities [39].

Graph-based attention mechanisms provide a complementary perspective, modeling feature relationships as graph structures to capture long-range dependencies. Graph Convolutional Networks (GCNs) [40] and their variants have been applied to object detection, enabling explicit reasoning about feature interactions. Hypergraph neural networks extend this concept by modeling higher-order correlations, which are particularly relevant for capturing complex structural relationships among objects in cluttered scenes [41]. However, traditional graph attention mechanisms suffer from two key limitations, typically employing dense connectivity patterns that scale poorly to high-resolution feature maps and lacking adaptive mechanisms to allocate computation according to scene complexity [42].

Our work addresses these limitations by introducing sparse hypergraph representations that capture higher-order feature relationships with low computational cost and by implementing dynamic sparsity adaptation based on local scene characteristics. This approach enables effective small object detection without the overhead of dense attention mechanisms. Compared with existing attention modules such as CBAM, ECA, coordinate attention, and global-context attention, the proposed DSHG module explicitly treats spatial positions as nodes and performs relational propagation through learnable hyperedge prototypes. Compared with dense graph or non-local reasoning, the proposed design avoids constructing an $N \times N$ pairwise affinity matrix and instead uses K prototypes, where $K \ll N$. Compared with dynamic convolution, which adapts convolutional kernels according to input features, DSHG adapts the residual contribution of prototype-mediated relational features according to a learned complexity score. Therefore, DASH-YOLO differs from a simple module combination by coupling complexity-aware relational reasoning with hierarchical attention refinement for small fitting detection.

3. Proposed Method

For clarity, the main tensor notations used in this section are defined as follows. A feature map is denoted by $X \in \mathbb{R}^{B \times C \times H \times W}$, where B, C, H , and W represent the batch size, number of channels, height, and width, respectively. The number of spatial positions is $N = H \times W$. Matrix multiplication is denoted by standard matrix product notation, T denotes transposition, $\odot$ denotes element-wise multiplication, and $\text{reshape}(\cdot)$ denotes tensor reshaping without changing the order of elements. When a scalar or vector coefficient is applied to a feature map, broadcasting follows the standard PyTorch convention. For example, the density coefficient $d \in \mathbb{R}^{B \times 1}$ is reshaped to $d \in \mathbb{R}^{B \times 1 \times 1 \times 1}$ and then broadcast along the channel and spatial dimensions during residual fusion.

3.1. Overall Architecture

DASH-YOLO derives from the YOLOv8x architecture and incorporates dedicated strategies to enhance deep feature extraction. The network consists of three key modifications. First, the backbone network integrates a Dynamic Sparse Hyper Graph mechanism to enhance deep feature representation. Second, a Hierarchical Attention module is applied immediately after the backbone to refine the deepest feature map. Third, the neck architecture employs adjusted C2f module configurations to improve feature recombination. Figure 1 illustrates the overall architecture. The following provides a detailed description of each component and its functionality. In the implemented architecture, the DSHG and HA modules are inserted after the final C2f block of the YOLOv8x backbone. For an input image of size 640×640 , the final backbone feature corresponds to the P5 level with a stride of 32, i.e., a spatial size of 20×20 and 1024 channels. The DSHG module is configured with 16 learnable hyperedge prototypes and a density range of $[0.3, 1.0]$. The output of DSHG is then refined by the HA module with a global attention downsampling ratio of 0.25. The neck follows the standard top-down and bottom-up multi-scale fusion path and outputs P3, P4, and P5 detection features.

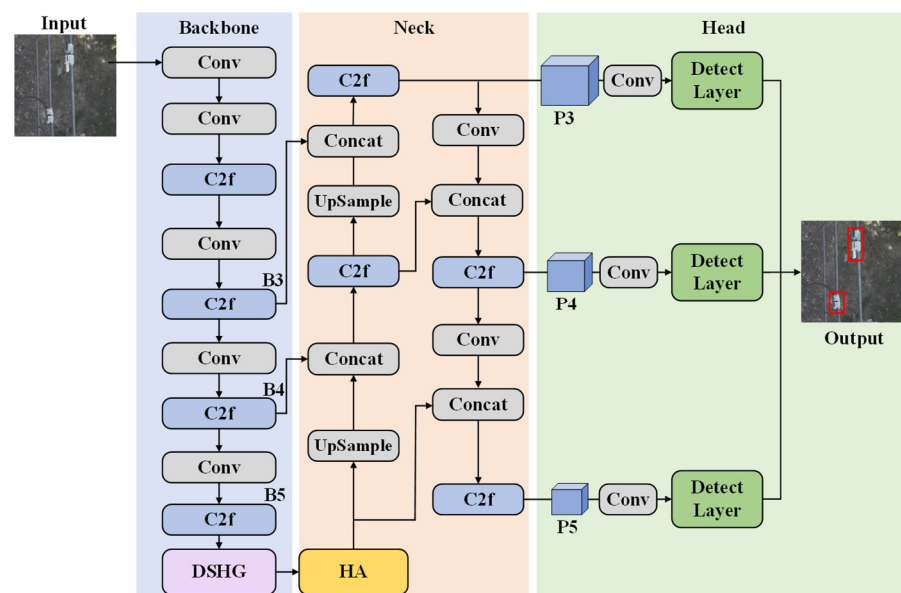


Figure 1. Overall architecture of DASH-YOLO.

The backbone integrates a Dynamic Sparse Hyper Graph mechanism at the deepest feature extraction stage to process the P5 feature map. This mechanism models complex feature relationships and long-range dependencies while maintaining computational efficiency through adaptive density control. By selectively enhancing informative feature interactions, it improves the representation of small and complex objects. The backbone outputs multi-scale features $\{P3, P4, P5\}$, providing complementary spatial and semantic information across different resolutions. The P5 feature is subsequently refined by a Hierarchical Attention module, which adaptively emphasizes discriminative regions and suppresses irrelevant responses, further strengthening feature representation. The neck architecture adopts C2f modules with disabled shortcut connections to promote effective feature recombination across scales. This design facilitates the integration of fine-grained spatial details with high-level semantic information while preserving high-resolution features. Through multi-scale fusion and feature propagation, the neck enhances robustness for object detection across varying object sizes and complex backgrounds.

3.2. Dynamic Sparse Hyper Graph Module

Traditional attention mechanisms in object detection suffer from quadratic computational complexity with respect to spatial resolution, which severely limits their applicability to high-resolution feature maps required for accurate small object detection. To overcome this limitation, we proposed the Dynamic Sparse Hypergraph (DSHG) module, which efficiently models long-range feature dependencies by constructing adaptive sparse hypergraph representations, thereby significantly reducing computational overhead.

The core motivation of the DSHG module lies in the observation that different spatial regions exhibit varying levels of structural complexity. Homogeneous background regions require only limited feature interaction, whereas regions containing small or densely distributed objects benefit from richer relational modeling. Instead of applying uniform attention across the entire feature map, the proposed DSHG module dynamically allocates computational resources according to regional complexity. As illustrated in Figure 2, the DSHG module comprises two synergistic components: a Complexity Predictor that evaluates regional feature complexity and a Simplified Hypergraph Interaction Layer that performs efficient feature propagation. Through an adaptive weighted residual fusion mechanism, these components generate the final enhanced feature map, where the dynamic weight modulates the contribution of hypergraph-enhanced features based on scene complexity. This adaptive density control strategy enables the module to dynamically regulate feature interaction strength according to regional characteristics, ensuring optimal feature enhancement while maintaining computational efficiency across diverse detection scenarios.

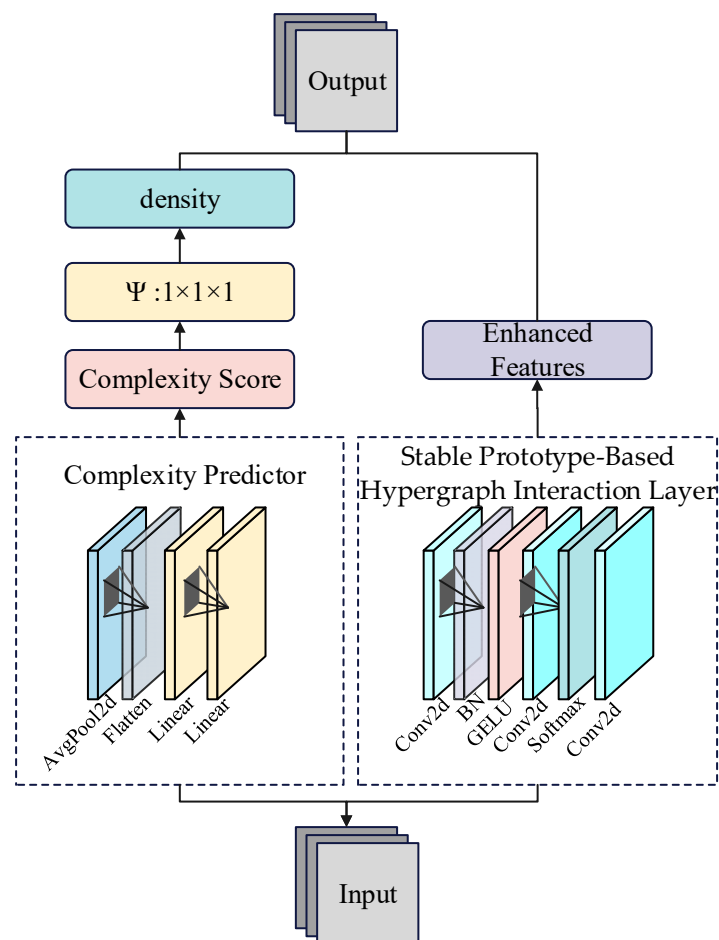


Figure 2. The overall architecture of Dynamic Sparse Hyper Graph.

The theoretical intuition for applying sparse hypergraph to small fitting detection is based on three properties of small object features in UAV inspection images. First, after backbone downsampling, each fitting is represented by only a few weak spatial activations, making isolated local evidence unreliable. Second, fittings are not distributed independently: clamps, rings, plates, conductors, and tower structures form repeated high-order co-occurrence patterns along transmission-line geometry. Third, most pixels belong to sky, vegetation, towers, or conductors, so dense pairwise attention over all spatial nodes introduces many redundant background interactions. The proposed hyper-graph representation addresses these properties by connecting multiple semantically related spatial nodes through a small set of learnable prototypes. A hyperedge can collect several weak responses from the same fitting or from structurally related fittings and then redistribute the aggregated context back to spatial nodes. Thus, DSHG enhances small-object features through higher-order context while avoiding the quadratic cost and background sensitivity of dense node-to-node interaction.

3.2.1. Complexity Predictor

The complexity predictor is designed as the core component of the adaptive computation framework, providing an efficient mechanism for scene complexity assessment. Given an input feature map $X \in \mathbb{R}^{B \times C \times H \times W}$, where B, C, H , and W denote the batch size, channel number, height, and width, respectively, the complexity predictor outputs a normalized complexity score $s \in \mathbb{R}^{B \times 1}$. The predictor first applies global average pooling to obtain a channel descriptor $g \in \mathbb{R}^{B \times C}$. The descriptor is then passed through two fully connected layers with a ReLU activation between them. The hidden dimension is set to $(C/4)$; for the P5 feature used in DASH-YOLO, ($C = 1024$), and the hidden dimension is therefore 256. Finally, a sigmoid function maps the output to the interval $[0, 1]$:

$$s = \sigma(W_2 \delta(W_1 g + b_1) + b_2) \quad (1)$$

where $W_1 \in \mathbb{R}^{C/4 \times C}$, $W_2 \in \mathbb{R}^{1 \times C/4}$, b_1 and b_2 are learnable biases, $\delta(\cdot)$ denotes ReLU, and $\sigma(\cdot)$ denotes the sigmoid function. A larger value of s indicates that the input feature contains richer structural information and therefore requires stronger hypergraph-enhanced residual refinement.

This predictor's lightweight nature enables seamless integration into deep architectures without introducing significant computational overhead, while providing reliable complexity priors that guide adaptive resource allocation throughout the network.

3.2.2. Stable Prototype-Based Hyper Graph Interaction Layer

To model long-range and higher-order feature relationships efficiently, the DSHG module employs a stable prototype-based hypergraph interaction layer. The input feature map $X \in \mathbb{R}^{B \times C \times H \times W}$ is first reshaped into a node sequence:

$$V = \text{reshape}(X) \in \mathbb{R}^{B \times N \times C}, \quad N = H \times W \quad (2)$$

where each node corresponds to one spatial position of the feature map. Layer normalization is then applied along the channel dimension to obtain $\hat{V}$. The layer maintains K learnable hyperedge prototypes ($P \in \mathbb{R}^{K \times C}$). In our implementation, $K = 16$, which is much smaller than the number of spatial nodes.

The first propagation stage aggregates node features into hyperedge features. The node-to-hyperedge affinity is computed as

$$A_{nh} = \text{Softmax}_N \left(\frac{\hat{V} P^T}{\sqrt{C}} \right) \quad (3)$$

where $A_{nh} \in \mathbb{R}^{B \times N \times K}$, and Softmax is performed along the node dimension for each hyperedge. The hyperedge representation is then obtained by weighted aggregation:

$$E = A_{nh}^T \hat{V}, \quad E \in \mathbb{R}^{B \times K \times C}$$

The second propagation stage sends information from hyperedges back to spatial nodes. The hyperedge-to-node affinity is calculated by

$$A_{hn} = \text{Softmax}_K \left(\frac{\hat{V} E^T}{\sqrt{C}} \right) \quad (4)$$

where $A_{hn} \in \mathbb{R}^{B \times N \times K}$, and Softmax is performed along the hyperedge dimension. The propagated node features are

$$V_p = A_{hn} E, \quad V_p \in \mathbb{R}^{B \times N \times C} \quad (5)$$

Finally, V_p is transformed by an MLP consisting of LayerNorm, a linear projection from C to $2C$, GELU activation, and a linear projection from $2C$ back to C . The output is reshaped to the original feature-map format:

$$H(X) = \text{reshape}(\text{MLP}(V_p)) \in \mathbb{R}^{B \times C \times H \times W} \quad (6)$$

Intuitively, each hyperedge prototype represents a latent interaction pattern shared by multiple spatial positions. Instead of comparing every node with every other node, the model first summarizes related nodes through a small set of prototypes and then redistributes the aggregated information back to all positions. This node-to-hyperedge-to-node route captures long-range dependencies while avoiding the quadratic cost of dense pairwise attention. The main interaction cost is proportional to $O(NKC)$, whereas dense pairwise attention requires $O(N^2C)$. For a P5 feature map of size 20×20 , $N = 400$ and $K = 16$, reducing the interaction scale from 400^2 pairwise relations to 400×16 prototype-mediated relations.

3.2.3. Dynamic Sparse Hyper Graph

The complexity score predicted by the complexity predictor is used to generate a density coefficient.

$$d = d_{\min} + (d_{\max} - d_{\min})s \quad (7)$$

where $d_{\min} = 0.3$ and $d_{\max} = 1.0$ in our implementation. Since $s \in \mathbb{R}^{B \times 1}$, the density coefficient d is broadcast to $\mathbb{R}^{B \times 1 \times 1 \times 1}$ during residual fusion. The final output of the DSHG module is formulated as

$$Y = X + d \cdot H(X) \quad (8)$$

where $H(X)$ denotes the hypergraph-enhanced feature generated by the prototype-based hypergraph interaction layer. The density coefficient does not change the tensor shape; it controls the strength of the hypergraph residual branch. When the predicted complexity is low, the module preserves more of the original feature and suppresses unnecessary relational enhancement. When the predicted complexity is high, the hypergraph-enhanced feature receives a larger residual weight, allowing the detector to emphasize long-range and higher-order dependencies in difficult scenes.

3.3. Hierarchical Attention Module

Standard attention mechanisms are unable to simultaneously capture global semantic context and preserve fine-grained spatial details, which limits their effectiveness for small

object detection. These miniature fittings occupy only a tiny fraction of the image area, yet their accurate detection requires both contextual understanding and pixel-level precision. To address this limitation, a Hierarchical Attention (HA) module is proposed, employing a coarse-to-fine refinement strategy that selectively enhances semantically relevant regions while suppressing irrelevant background features.

Within the HA module, a Global Attention Net identifies contextually significant regions through hierarchical downsampling and upsampling, selectively retaining the most informative portions of the feature map, whereas a Local Attention Net performs pixel-level refinement using depth-wise separable convolution to preserve fine spatial details. The outputs of these two branches are integrated via a residual fusion mechanism, producing attention enhanced features that combine contextual awareness with spatial precision. The overall architecture of the HA module is illustrated in Figure 3, highlighting the interaction between the global and local attention pathways. This design substantially reduces computational overhead compared to full-resolution attention while maintaining critical spatial details required for small fitting detection.

For reproducibility, the input of the HA module is denoted as $X_h \in \mathbb{R}^{B \times C \times H \times W}$. The global branch first downsamples X_h by a scale factor of 0.25 using bilinear interpolation. The downsampled feature is processed by two 1×1 convolutions, where the channel dimension is reduced from C to $C/8$ and then projected to a single-channel attention map. After sigmoid activation, the attention map is upsampled back to $H \times W$ and multiplied with the input feature. The local branch then applies a 3×3 depth-wise convolution followed by sigmoid activation to generate spatial modulation weights at full resolution. The final HA output is obtained by residual fusion:

$$Y_h = X_h + \text{LocalAttn}(X_h \odot A_g) \quad (9)$$

where $A_g \in \mathbb{R}^{B \times 1 \times H \times W}$ is the global attention map, and $\odot$ denotes element-wise multiplication. This implementation uses continuous attention modulation rather than hard region selection, which keeps the operation differentiable and stable during training.

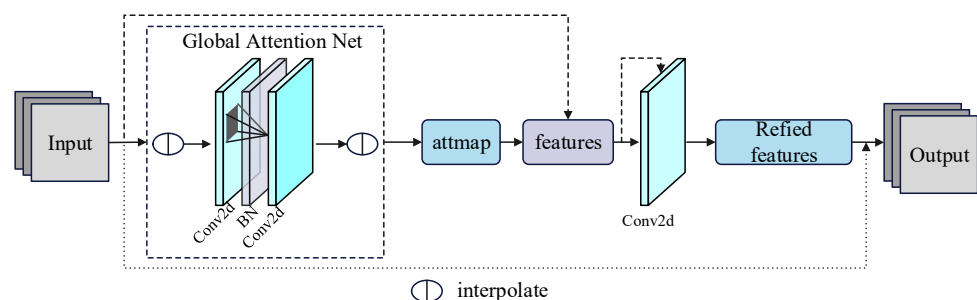


Figure 3. The overall architecture of Hierarchical Attention module.

3.3.1. Global Attention Net

The Global Attention Net is designed to efficiently model global contextual information while avoiding the high computational cost associated with direct high resolution processing. Instead of operating on the original feature resolution, the proposed module adopts a downsample and upsample architecture, which enables effective global context aggregation with significantly reduced computational complexity.

The input of the Hierarchical Attention module is the output feature map from the DSHG module in Figure 1. For notational clarity in the following analysis, this input

feature map is denoted as $Y \in \mathbb{R}^{C \times H \times W}$, the spatial resolution is first reduced via bilinear downsampling with a fixed scale factor $s = 0.25$, yielding

$$Y_1 = \text{Downsample}(Y; s) \quad (10)$$

where the downsampling ratio is empirically selected to balance computational efficiency and semantic information preservation. This spatial compression reduces the feature map resolution while retaining critical contextual cues necessary for global attention modeling. The compressed feature map is then processed through a two-stage transformation pipeline.

First, a 1×1 convolution with a weight matrix $W_1 \in \mathbb{R}^{(C/8) \times C \times 1 \times 1}$ is applied to reduce channel dimensionality while preserving semantic capacity, resulting in

$$Y_2 = \text{Conv}_{1 \times 1}(Y_1, W_1) \quad (11)$$

Second, another 1×1 convolution generates a single channel attention representation $M_1 \in \mathbb{R}^{1 \times (sH) \times (sW)}$, followed by a sigmoid activation:

$$M_1 = \sigma(\text{Conv}_{1 \times 1}(Y_2)) \quad (12)$$

where $\sigma(\cdot)$ denotes the sigmoid activation function. The resulting attention map is subsequently upsampled to the original spatial resolution via bilinear interpolation:

$$M_g = \text{Upsample}(M_1, H \times W) \quad (13)$$

where $H \times W$ denotes the spatial resolution of Y . Yielding a smooth and spatially consistent global guidance map $M_g \in \mathbb{R}^{1 \times H \times W}$.

Through this hierarchical design, global contextual information is captured at a reduced spatial resolution, leading to substantial computational savings compared to direct high-resolution attention mechanisms. Meanwhile, the use of fully differentiable operations ensures stable gradient propagation during training. The resulting global guidance map effectively highlights semantically significant regions, providing reliable global context cues for subsequent feature enhancement stages.

3.3.2. Local Attention Net

The Local Attention Net implements fine grained spatial feature refinement through an efficient depth-wise separable convolution mechanism. Given an input feature map $X \in \mathbb{R}^{C \times H \times W}$, the module applies a 3×3 depth-wise separable convolution followed by a sigmoid activation to generate spatially adaptive modulation weights:

$$A = \sigma(\text{DepthwiseConv}_{3 \times 3}(Y)) \quad (14)$$

Compared with standard convolution, this operation significantly reduces computational complexity from $O(C \cdot K^2)$ to $O(K^2)$, where K denotes the kernel size. For example, when $C = 1024$ and $K = 3$, the relative computation cost is reduced to $9/(1024 + 9) \approx 0.0088$, corresponding to approximately 99.1% computational savings, while preserving full spatial resolution.

The resulting spatial attention weights $A \in \mathbb{R}^{C \times H \times W}$ precisely model the semantic importance distribution across all spatial positions. The module completes the refinement process through element wise multiplication:

$$M_l(Y) = Y \odot A \quad (15)$$

This operation enhances discriminative local responses while suppressing background noise, without altering the spatial structure of the feature map. Owing to the channel-wise independence of depth-wise convolution, fine spatial details are preserved throughout the refinement process, particularly critical for small object localization, enabling precise local enhancement with minimal additional computational overhead.

3.3.3. Hierarchical Attention

The Hierarchical Attention (HA) module integrates global semantic guidance and local spatial refinement into a unified feature enhancement framework through a two-stage coarse-to-fine cascade. The module first employs the global guidance map M_g to screen the input features:

$$Y_i = Y \odot M_g, Y_i \in \mathbb{R}^{C \times H \times W} \quad (16)$$

This global screening process selectively retains semantically informative regions, while effectively filtering out redundant information. The filtered features are then refined through fine grained local attention:

$$Y_r = M_l(Y_i) \quad (17)$$

The coarse to fine processing strategy avoids full-image refinement, enabling efficient allocation of computational resources while enhancing feature discriminability in regions critical for small object detection.

The final output is obtained via a residual fusion strategy:

$$Z = Y + Y_r \quad (18)$$

The residual design ensures stable optimization and preserves original feature information, allowing the module to gracefully degenerate to an identity mapping when refinement is unnecessary. By combining global semantic filtering with localized spatial enhancement, the proposed HA module achieves an effective balance between computational efficiency and representation capacity, particularly benefiting dense scenes and small object detection scenarios without introducing substantial computational overhead.

3.4. Training Strategy and Loss Function

The training objective adopts the standard YOLO multi-component loss function that jointly optimizes localization accuracy and classification performance. Loss design is especially relevant for small-object detection because tiny localization errors can cause large IoU fluctuations, and long-tailed categories may be under-optimized. Prior work has addressed these issues using focal loss to reduce the dominance of easy negatives, IoU-based localization losses to improve bounding-box regression, and distribution focal loss for more precise box boundary modeling. In this study, however, no additional small-object-specific loss term, class-dependent weight, or focal-loss modification were introduced. This design choice keeps the comparison focused on the proposed architectural changes, namely DSHG and HA, rather than mixing architectural gains with loss reweighting.

$$L_{total} = \lambda_{box} \cdot L_{box} + \lambda_{cls} \cdot L_{cls} + \lambda_{dfl} \cdot L_{dfl} \quad (19)$$

where λ_{box} , λ_{cls} , and λ_{dfl} denote the loss gains for bounding-box regression, classification, and distribution focal loss, respectively. Unless otherwise stated, we used the standard YOLO loss gains $\lambda_{box} = 7.5$, $\lambda_{cls} = 0.5$, and $\lambda_{dfl} = 1.5$. No additional small-object-specific loss term or custom class-dependent weighting was introduced, so that the

performance gains can be attributed to the proposed architectural modifications rather than to loss reweighting.

To improve small object detection performance, a set of targeted data augmentation strategies was employed. Mosaic augmentation was used to increase small object density and contextual diversity through multi-image composition, while copy–paste augmentation explicitly replicated small instances to enhance instance-level diversity. In addition, adaptive RandAugment was applied with photometric distortions and horizontal flipping to improve robustness against appearance variations. To avoid excessive perturbations in later training stages, mosaic augmentation was disabled after the early epochs, preserving early gains in small-object representation while reducing the risk of overfitting.

The model was trained for 1000 epochs using the AdamW optimizer, with an initial learning rate of 0.01, linearly decayed to 1×10^{-4} , and a short warmup phase to stabilize early optimization. A weight decay of 5×10^{-4} was applied for regularization. All experiments were conducted with a fixed input resolution of 640 pixels.

4. Experimental Evaluation

4.1. Dataset Construction

A dedicated dataset for transmission-line fitting detection was constructed to address the lack of publicly available benchmarks for this specific inspection scenario. The dataset comprises 820 high-resolution UAV images with a spatial resolution of 4000×3000 pixels. The images were captured using a DJI Phantom 4 Pro UAV (SZ DJI Technology Co., Ltd., Shenzhen, China) during routine transmission-line inspection missions across multiple regions of China, encompassing diverse terrains such as mountainous areas, urban regions, forests, and plains. To ensure the representativeness of real-world inspection scenarios, the acquisition process covered extensive variations in illumination, season, weather, inspection distance, and background complexity, thereby comprehensively reflecting practical UAV inspection conditions.

The dataset encompassed 15,929 fitting instances corresponding to ten common types of transmission line fittings, functionally categorized into three groups: clamps, dampers, and connectors. Clamps included parallel groove, grounding, suspension, and strain clamps; dampers consisted solely of anti-vibration hammers; and connectors covered U-type hanging rings, adjusting plates, ball head hanging rings, bowl head hanging plates, and right-angle hanging plates. In terms of class distribution, connectors accounted for 52.9% of the samples, followed by clamps at 44.7%, while dampers represented only 2.4%. This distribution was consistent with real-world inspection scenarios, where dampers are sparsely deployed at specific locations, whereas clamps and connectors are extensively used as core load-bearing and linking components. Table 1 further provides the exact instance counts for all ten fitting categories, highlighting the pronounced long-tailed imbalance at the fine-grained class level. Such fine-grained imbalance is important for interpreting the validation results, because rare categories and visually similar connector classes may be more susceptible to missed detections or inter-class confusion during model evaluation.

Table 1. Exact instance counts of the ten fitting categories.

ID	Fitting Type	Instances	Avg.Size(px ²)
1	Anti-vibration Hammer	393	15,595.73
2	Parallel Groove Clamp	1309	4636.06
3	Grounding Clamp	676	30,304.17
4	Suspension Clamp	3888	10,109.26

Table 1. Cont.

ID	Fitting Type	Instances	Avg.Size(px ²)
5	Strain Clamp	1244	56,791.44
6	U-type Hanging Ring	651	9596.03
7	Adjusting Plate	698	79,492.93
8	Ball Head Hanging Ring	2393	3295.17
9	Bowl Head Hanging Plate	2277	7631.04
10	Right Angle Hanging Plate	2400	11,328.9

All images were manually annotated with bounding boxes and category labels under strict quality-control procedures. The annotation guideline required each bounding box to tightly enclose the visible region of a fitting while excluding adjacent conductors, insulators, and background structures as much as possible. For partially occluded fittings, annotators labeled the visible object extent when the category could be identified with confidence; severely blurred or ambiguous objects were flagged for expert review. The labels were exported from CVAT and subsequently checked by experienced power-grid inspection engineers. Disagreements in object extent or category assignment were discussed and corrected through consensus review before the final annotation files were used for training and evaluation. As an additional consistency check between the final manual annotations and the trained detector, predictions generated by the best DASH-YOLO checkpoint on the validation set were matched to the ground-truth boxes using an IoU threshold of 0.5. The matched bounding boxes achieved a mean IoU of 0.8729, and the category assignments reached a Cohen's Kappa coefficient of 0.9989, indicating highly consistent localization and category prediction for matched objects. Based on these annotations, Figure 4 illustrated the normalized width and height distribution of transmission line fittings, revealing pronounced scale sparsity with the majority of targets densely clustered in the small-size region. This distribution indicates that most fittings occupy only a very small proportion of the image area. Such extreme small-object characteristics directly motivated the adoption of the proposed Hierarchical Attention mechanism, which is specifically designed to preserve and enhance micro-scale features.

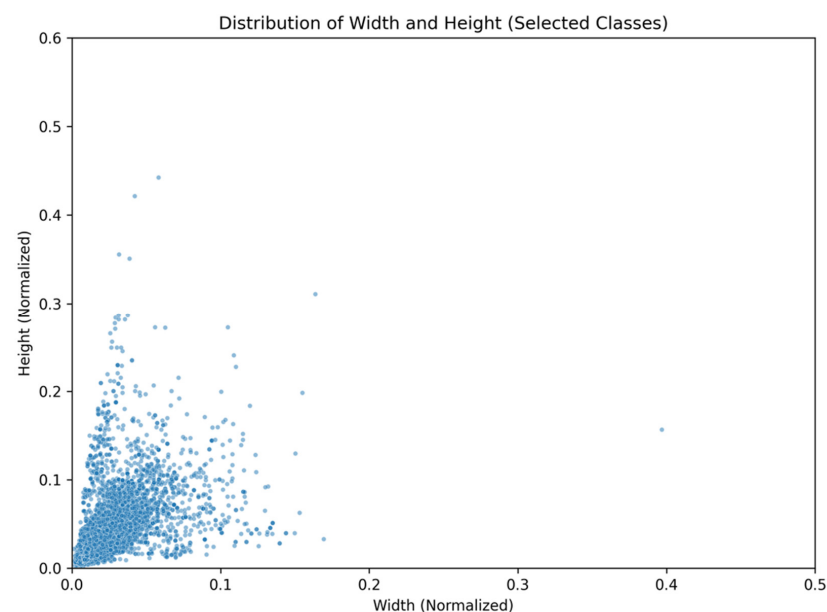


Figure 4. Distribution of width and height of transmission line fittings.

For model training and validation, the dataset was split at the image level into training and validation subsets at an 80:20 ratio using a fixed random seed. Because each image can contain multiple fitting categories, the split was checked against the category distribution to avoid severe category imbalance between subsets. The experimental results are based on this fixed training-validation split. Although k-fold cross-validation could provide a more comprehensive assessment of model robustness, it was beyond the scope of the present study and will be considered in future work.

Data availability: Due to power grid security and privacy constraints, full dataset cannot be fully public, and the dataset is available on request from the authors after signing confidentiality agreements.

4.2. Experimental Setup

All experiments were conducted on a workstation equipped with an NVIDIA RTX 3090 GPU, an Intel i7-12700F processor, and 32 GB of RAM. The models were implemented using PyTorch 2.2 with CUDA 11.8 and cuDNN 8.7.0 to enable GPU acceleration.

For fair comparison, all models were trained for 1000 epochs using the AdamW optimizer with an initial learning rate of 0.01, which was linearly decayed to 0.0001. A weight decay of 5×10^{-4} and a momentum of 0.937 were applied. Training was performed with a batch size of 32 using the default Ultralytics random seed (seed = 0) with deterministic mode enabled. To improve training stability and data diversity, warmup, automatic mixed-precision training, adaptive RandAugment with photometric distortions, horizontal flipping, and validation-based best-checkpoint selection were adopted. The same train/validation split, input resolution, optimization settings, data augmentation strategy, and evaluation protocol were used for all compared methods.

The original UAV images have a resolution of 4000×3000 pixels, whereas the training and validation experiments in this paper use the standard 640×640 YOLO input. During training and validation, images are resized with aspect-ratio-preserving letterbox preprocessing. For deployment on raw high-resolution images, direct resizing of the whole image may further shrink very small fittings. Therefore, practical inference should use sliding-window or tiled prediction with overlap, followed by non-maximum suppression or weighted box fusion across tile boundaries. The FPS presented in Table 2 corresponds to a single 640×640 input on the stated workstation; end-to-end latency on raw 4000×3000 images depends on the number of tiles and image-overlap setting.

Table 2. Performance comparison with state-of-the-art methods.

Method	mAP@0.5	mAP@0.5:0.95	Recall	Precision	Params (MB)	FLOPs (G)	FPS (f/s)
Faster R-CNN	74.55	60.32	71.82	75.17	137.12	370.2	49.65
DETR	70.38	56.14	68.55	69.36	76.31	142.7	58.31
YOLOv8x	85.74	70.12	85.91	79.67	68.18	129.14	57.59
YOLOv10x	86.05	66.06	85.86	77.44	61.30	85.65	61.87
YOLOv13x	89.00	65.06	88.68	82.23	64.0	118.28	59.17
DASH-YOLO	90.37	74.74	90.80	85.49	65.73	95.73	62.99

4.3. Evaluation Metrics

To comprehensively evaluate detection accuracy and computational efficiency, standard object detection metrics are adopted together with efficiency-related indicators, with particular emphasis on small-object detection performance.

Detection performance is primarily assessed using mean Average Precision at an IoU threshold of 0.5 and mean Average Precision averaged over IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05, which are calculated as

$$mAP@0.5 = \frac{1}{n} \sum_{i=1}^n AP_i(IoU \geq 0.5) \quad (20)$$

$$mAP@0.5:0.95 = \frac{1}{10n} \sum_{k=1}^{10} \sum_{i=1}^n AP_i^{(k)} \quad (21)$$

where n denotes the number of object categories and AP_i represents the Average Precision of the i -th category. $AP_i^{(k)}$ denotes the Average Precision of category at the i -th IoU threshold.

Precision and Recall are also employed to measure detection reliability and completeness, respectively. They are defined as:

$$Precision(P) = \frac{TP}{TP + FP} \quad (22)$$

$$Recall(R) = \frac{TP}{TP + FN} \quad (23)$$

where TP , FP , and FN denote the numbers of true positives, false positives, and false negatives.

In addition to accuracy-related metrics, computational efficiency is evaluated using Frames Per Second (FPS), which measures the number of images processed per second during inference and reflects real-time capability. And the computational complexity of each model is measured in terms of FLOPs and the number of parameters, where FLOPs indicate the total number of floating-point operations required for a single forward pass under a fixed input resolution.

4.4. Comparison with State-of-the-Art Methods

To evaluate the effectiveness of DASH-YOLO, we conducted a comprehensive comparison with five state-of-the-art object detection methods, including two-stage detectors Faster R-CNN, the transformer-based DETR, and YOLO family variants with YOLOv8x, YOLOv10x, YOLOv13x. All baseline models were trained under identical data splits and augmentation strategies to ensure fair comparison. The quantitative comparison results of different methods are summarized in Table 2.

In terms of detection performance, DASH-YOLO consistently outperforms all five compared methods across the main evaluation metrics. Specifically, DASH-YOLO achieves an improvement of 4.63 percentage points in $mAP@0.5$ over YOLOv8x, which serves as a strong baseline, and still exceeds the best-performing existing model YOLOv13x by 1.37 percentage points. Under the stricter $mAP@0.5:0.95$ metric, DASH-YOLO demonstrates even more pronounced gains, achieving improvements of 4.62 percentage points over YOLOv8x and more than 8 percentage points over YOLOv13x, indicating substantially enhanced localization accuracy. Compared with the remaining two-stage and transformer-based detectors, DASH-YOLO yields even larger margins, confirming its robustness for small and complex object detection in UAV imagery. In addition, DASH-YOLO achieves the highest recall and precision among all evaluated methods, reflecting a balanced detection behavior that effectively reduces both missed detections and false positives.

From the perspective of computational efficiency, DASH-YOLO improves the trade-off rather than minimizing model size alone. Compared with YOLOv8x, DASH-YOLO reduces FLOPs from 129.14 G to 95.73 G, corresponding to an approximately 25.9% reduction under the 640×640 input setting. The parameter size remains comparable, decreasing

slightly from 68.18 MB to 65.73 MB. Therefore, the efficiency improvement should be interpreted mainly as a moderate FLOPs reduction and an inference-speed gain, not as a drastic reduction in the total number of parameters. This reduction is primarily attributed to the proposed adaptive sparse computation strategy, which prioritizes effective feature interaction over redundant dense computation. Under the same testing environment, DASH-YOLO achieves 62.99 FPS, compared with 57.59 FPS for YOLOv8x and 59.17 FPS for YOLOv13x, demonstrating a more favorable accuracy and efficiency trade-off than existing state-of-the-art detectors.

The experimental results indicate that DASH-YOLO provides a clear validation-set advantage over the evaluated YOLO variants and representative detectors from other paradigms under the same training protocol. The method performs well in small fitting scenarios while maintaining real-time inference at the 640×640 input resolution. Broader claims about field deployment require additional multi-seed, cross-dataset, and raw-image tiled-inference evaluation.

4.5. Statistical Reliability and Validation

Under the fixed-seed training setup described in Section 4.2, the convergence behavior was also monitored to assess optimization stability. As shown in Figure 5, the training box, classification, and DFL losses decreased rapidly in the early epochs and then converged gradually. The validation box and classification losses followed similar decreasing trends, while precision, recall, mAP@0.5, mAP@0.75, and mAP@0.5:0.95 increased steadily and approached stable plateaus. These curves indicate stable convergence and no obvious late-stage divergence between training and validation behavior, suggesting that the experimental validation-set improvements arise from a stable optimization process rather than training fluctuation.

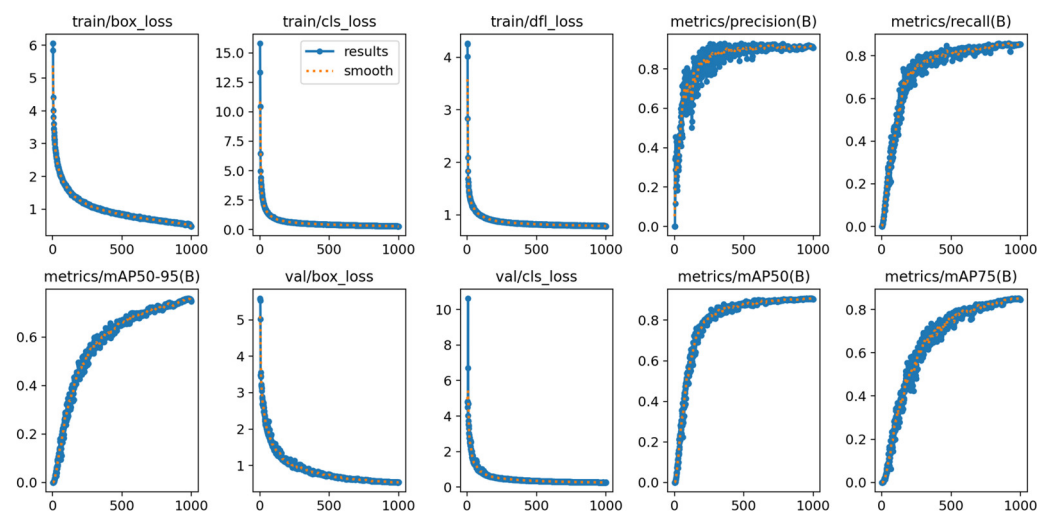


Figure 5. Training and validation curves of DASH-YOLO over 1000 epochs.

The current evidence should be interpreted within this validation scope. Because the study presents a single fixed-seed run on a fixed training/validation split, it does not provide multi-seed standard deviations, confidence intervals, or independent test-set significance estimates. In addition, public UAV or small-object datasets such as VisDrone and COCO-small lack the same fine-grained transmission-line fitting categories, so direct category-level external validation is not fully comparable. Therefore, validation focuses on the proposed fitting dataset and its label space, while broader cross-dataset reliability remains future work.

4.6. Ablation Studies

To further investigate the contribution of each proposed module and clarify their respective roles in performance improvement, we conducted a series of ablation experiments based on YOLOv8x. All variants were trained and evaluated under identical settings to ensure fair comparison. Table 3 summarized the quantitative results, based on which the effects of the Dynamic Sparse Hyper Graph (DSHG) module and the Hierarchical Attention (HA) module are analyzed separately.

Table 3. Ablation Study of Different Module.

Method	mAP@0.5	mAP@0.5:0.95	Recall	Precision	Params (MB)	FLOPs (G)	FPS (f/s)
yolov8x	85.74	70.12	85.91	79.67	68.18	129.14	57.59
yolov8x-DSHG	90.07	72.88	92.18	85.59	65.45	95.7	59.96
yolov8x-HA	89.66	72.12	90.80	84.59	64.10	95.26	60.47
DASH-YOLO	90.37	74.74	93.32	85.49	65.73	95.73	62.99

4.6.1. Ablation Study of the DSHG Module

Introducing the Dynamic Sparse Hyper Graph (DSHG) module into the YOLOv8x baseline leads to a substantial performance improvement across multiple evaluation metrics. Compared with the baseline model, YOLOv8x-DSHG achieves a notable increase in mAP@0.5 from 85.74 to 90.07, accompanied by a clear enhancement in precision, which rises from 79.67 to 85.59. This improvement indicates that the DSHG module is particularly effective in refining feature representations and suppressing false positives by modeling long-range and higher-order feature dependencies.

In terms of recall, the DSHG-enhanced model also exhibits a significant gain, increasing from 85.91 to 92.18, demonstrating that richer contextual interactions help the detector capture more true targets, especially small and sparsely distributed objects. Meanwhile, the introduction of DSHG does not impose additional computational burden. On the contrary, both parameter size and FLOPs are reduced compared with the baseline, while inference speed increases from 57.59 to 59.96 FPS. This suggests that the dynamic sparsification mechanism enables more efficient feature interaction by selectively preserving informative relations and discarding redundant computations.

Overall, these results confirm that the DSHG module not only improves detection accuracy but also enhances computational efficiency, achieving a favorable balance between effectiveness and efficiency through adaptive feature interaction modeling.

4.6.2. Ablation Study of the HA Module

When only the Hierarchical Attention module is integrated into the baseline, the resulting YOLOv8x-HA variant also demonstrates consistent performance gains. Specifically, mAP@0.5 increases to 89.66, while recall reaches the highest value among all single-module variants at 93.32. This pronounced recall improvement highlights the HA module's strength in multi-scale feature fusion and progressive attention allocation, which facilitates more comprehensive object coverage across different spatial resolutions.

Although the precision achieved by YOLOv8x-HA is slightly lower than that of the DSHG-only variant, it still represents a clear improvement over the baseline, indicating that hierarchical attention effectively enhances feature discrimination without introducing excessive false detections. From an efficiency perspective, the HA module yields the lowest parameter count and FLOPs among all configurations, while further boosting inference speed to 60.47 FPS. This demonstrates that the hierarchical design efficiently reuses and refines multi-level features rather than increasing computational complexity.

These observations suggest that the HA module primarily contributes to improving detection completeness and real-time performance, making it especially beneficial for scenarios where high recall and fast inference are critical.

The ablation results indicate that DSHG and HA contribute complementary effects rather than redundant improvements. The DSHG-only variant improves precision and mAP by strengthening deep relational representations through prototype-mediated hypergraph propagation. The HA-only variant improves recall by combining global semantic guidance with local spatial refinement, which helps recover weak small-object responses. The full DASH-YOLO model achieves the best mAP@0.5:0.95 and the highest overall recall, showing that the two modules act at different stages of feature enhancement. DSHG first improves the semantic consistency of the deepest feature map, while HA further refines spatially discriminative responses before multi-scale fusion. This staged design explains why the full model performs better than either single-module variant under the same training and evaluation protocol.

4.6.3. Hyper-Parameter Sensitivity Analysis

To improve transparency regarding the main hyper-parameters of the proposed DSHG module, we provide an implementation-level sensitivity and complexity analysis. In the implemented model, DSHG is inserted at the P5 stage with $C = 1024$, $H = W = 20$, $N = H \times W = 400$, $K = 16$, and $[d_{\min}, d_{\max}] = [0.3, 1.0]$. The number of hyperedge prototypes K directly controls the capacity and cost of prototype-mediated relational propagation. Since the node-to-hyperedge and hyperedge-to-node affinities both have size $N \times K$, the interaction scale grows linearly with K , i.e., $O(NKC)$, rather than quadratically with N . The density range controls only the residual strength $d = d_{\min} + (d_{\max} - d_{\min})s$, and therefore changes the feature modulation intensity without changing the tensor shape, parameter count, or FLOPs of the hypergraph branch. The compression ratio of the complexity predictor controls the hidden dimension of the two-layer predictor. In the current setting, the predictor uses a $C/4$ hidden dimension, which corresponds to 262,657 learnable parameters when $C = 1024$. These implementation-level effects are summarized in Table 4.

Table 4. Implementation-level sensitivity and complexity analysis of DSHG hyper-parameters.

Hyper-Parameter Category	Setting	Main Implementation Effect	Affected Parameters	Interaction Scale on P5 (N = 400)
Number of Prototypes (K)	K = 8	Fewer prototypes; lower relational capacity and interaction cost	8192	$N \times K = 3200$ affinities
	K = 16 (Default)	Balanced relational capacity and computational cost	16,384	$N \times K = 6400$ affinities
	K = 32	More prototypes; higher relational capacity and interaction cost	32,768	$N \times K = 12,800$ affinities
Residual Modulation Range ($d_{\min}, d_{\max}$)	[0.1,1.0]	Wider modulation range; weaker lower-bound enhancement for simple scenes	0 (no change)	No change
	[0.3,1.0]	Preserves moderate minimum hypergraph residual contribution	0 (no change)	No change
	[0.5,1.0]	Stronger minimum residual contribution; less suppression in simple scenes	0 (no change)	No change
Complexity Predictor Compression Ratio	Ratio = 1/2	Hidden dimension = 512; larger predictor capacity	525,313	No change
	Ratio = 1/4 (Default)	Hidden dimension = 256; balanced predictor size and accuracy	262,657	No change
	Ratio = 1/8	Hidden dimension = 128; more compact predictor	131,329	No change

4.7. Category-Level Difficulty Analysis

Although per-category AP, precision, and recall statistics are not provided in this study, a category-level difficulty analysis is presented based on the verified dataset statistics in Table 1, including the instance counts and average bounding-box areas, together with qualitative detection examples. Across repeated validation observations, the categories with lower detection performance were consistently concentrated among small, sparse, or visually similar fittings. Ball Head Hanging Ring and Parallel Groove Clamp have the smallest average target areas, with average bounding-box areas of 3295.17 px^2 and 4636.06 px^2 , respectively, making them more vulnerable to feature loss after downsampling. Anti-vibration Hammer has only 393 instances, representing the most data-scarce category and thus a higher risk of missed detection under limited training samples. U-type Hanging Ring and Bowl Head Hanging Plate are also challenging because they have relatively small target sizes and visual similarity to adjacent connector categories. In contrast, larger categories such as Adjusting Plate and Strain Clamp usually provide stronger spatial cues, although they may still be confused under occlusion or cluttered backgrounds. This analysis indicates that the relatively difficult fittings are mainly those with small physical image size, limited sample frequency, and high inter-class visual similarity.

5. Visualization Analysis

To visually demonstrate the detection capabilities of DASH-YOLO and the internal operation mechanism of the proposed module, a comprehensive visual analysis is provided, including feature visualizations, attention maps and detection results under various situations.

5.1. Feature Map Visualization

To further understand how the DSHG module preserves feature integrity throughout the network hierarchy, we analyzed the internal feature representations at different depth levels. Figure 6 visualizes feature map activations at layer 14 for baseline YOLOv8x and DASH-YOLO, comparing how different feature channels respond to small object targets. Layer 14 represents a critical fusion stage in the neck module, where channel 0 encodes edge-based structural features, channel 2 captures mid-level texture patterns, and channel 4 represents high-level semantic information for small object categories.

The comparison revealed distinct differences in activation patterns across network depths. Both models exhibited comparable responses in shallow layers to low-level features such as edges and textures. However, in deeper layers where small object features typically diminish due to hierarchical downsampling, the proposed algorithm with the DSHG module demonstrated significantly stronger and more localized activations. It produced concentrated high-intensity regions around small targets such as transmission line fittings, with clear focus on the target areas. In contrast, the baseline exhibited weaker and more diffuse activations, particularly in channels 2 and 4 where features appeared less discriminative and scattered. Channel-wise analysis revealed consistent advantages for DSHG across all three channels. Channel 0 showed enhanced edge feature preservation with pronounced high-activation regions at object boundaries. Channel 2 exhibited improved texture pattern concentration, with activation intensities focusing around target regions rather than dispersing across the feature map. Channel 4 displayed the most striking difference, with multiple distinct high-intensity clusters precisely aligned with small object locations.

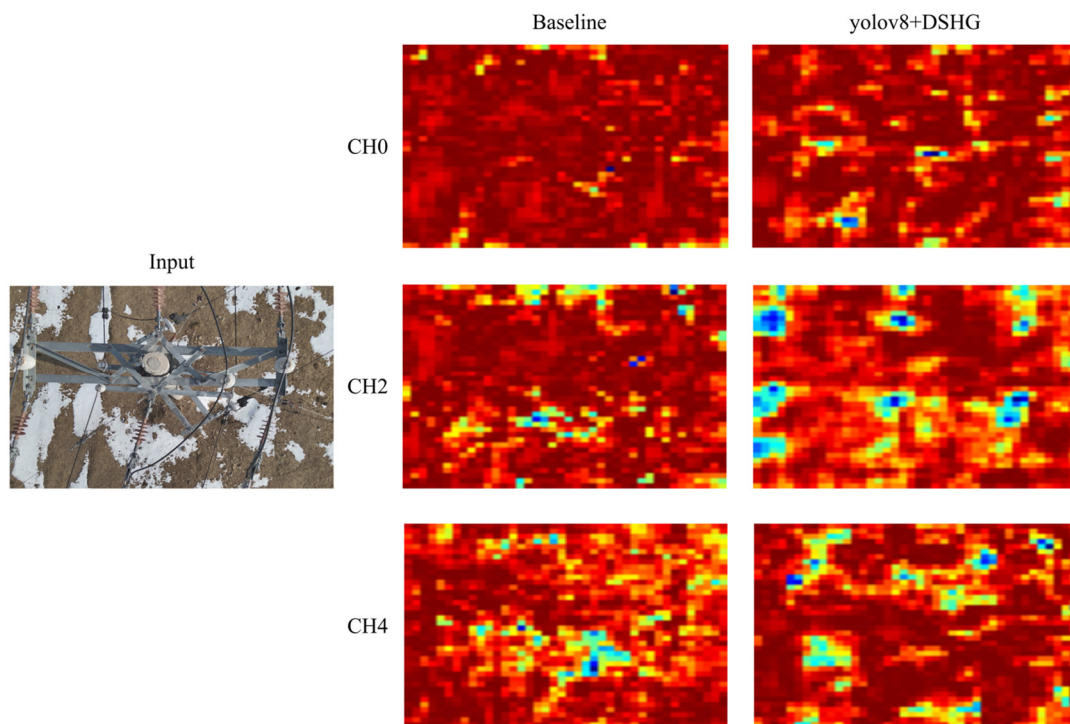


Figure 6. Comparison of feature maps between the baseline of the 14th layer network output channels 0, 2, and 4 and the YOLOv8x + DSHG.

Mechanistically, the DSHG module enhances the preservation of these discriminative features through adaptive sparse hypergraph interactions. This mechanism dynamically prioritizes informative regions corresponding to small objects while suppressing irrelevant background noise, effectively mitigating feature degradation in deeper layers without introducing additional computational overhead. Collectively, these visual results confirm that the Dynamic Sparse Hyper Graph module effectively maintains feature integrity for small object detection, providing qualitative evidence for the superior performance of DASH-YOLO over the baseline YOLOv8x.

5.2. Attention Map Visualization

Among the various visualization strategies, attention map visualization was employed to analyze where the model allocates its focus during detection, thereby providing insights into the feature selection behavior induced by the proposed module. Figure 7 presented qualitative comparisons of Grad-CAM attention maps between the baseline YOLOv8x and the Hierarchical Attention (HA) enhanced variant under diverse aerial inspection scenarios.

As shown in Figure 7a, in scenes containing sparsely distributed small fittings, the HA-enhanced model produced more compact and target-focused activation patterns, whereas the baseline model exhibited scattered responses across surrounding background regions. This indicated that HA improved attention concentration on small and spatially isolated components. In complex background scenes with partial occlusions (Figure 7b), the HA-enhanced variant maintained more continuous and coherent attention responses over the visible parts of the fittings, while the baseline attention maps showed fragmented and unstable activations influenced by background structures. Similar behavior was observed under low-light conditions (Figure 7c), where HA preserved structured attention patterns aligned with salient object regions, whereas the baseline model suffered from increased noise and attention diffusion. For dense target scenarios Figure 7d, the hierarchical attention mechanism demonstrated improved discrimination between closely spaced fittings by distributing attention across multiple relevant regions without excessive background inter-

ference. This behavior reflected the multi-scale nature of HA, in which global contextual cues and local spatial details were jointly leveraged to enhance feature representation.

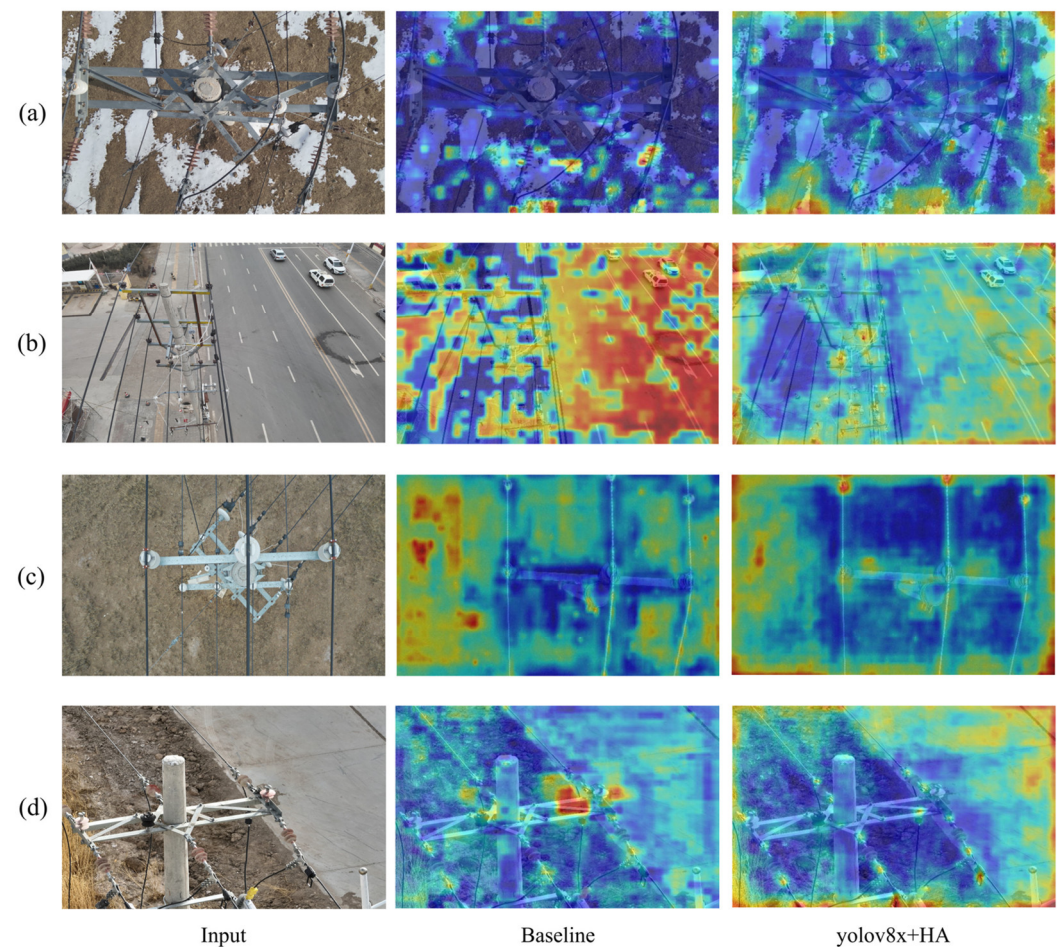


Figure 7. Comparison of Grad-CAM heat maps in aerial images under various situations. (a) Sparse fittings. (b) Complex backgrounds (c) Low-visibility conditions. (d) Dense fittings.

Collectively, the visual comparisons consistently showed that the proposed HA mechanism enhanced attention stability, spatial coherence, and target relevance across challenging inspection conditions. These qualitative results are consistent with the quantitative improvements presented above, further suggesting that the HA module contributes to stronger and more discriminative feature representation for small and visually complex transmission-line fittings.

5.3. Detection Results Visualization

To further validate the practical effectiveness of DASH-YOLO in real-world transmission line inspection scenarios, Figure 8 provided comprehensive qualitative detection comparisons between DASH-YOLO and baseline YOLOv8x in four representative challenging scenarios, including scenes with densely distributed targets, sparsely distributed targets, low-light environments, and complex backgrounds.

In dense scenes (Figure 8a) with multiple small fittings clustered along transmission lines, DASH-YOLO detected more annotated instances with tighter localization than YOLOv8x in the illustrated examples, including several partially occluded or background-embedded targets. For sparse distribution scenarios (Figure 8b), DASH-YOLO showed more stable bounding-box coverage, whereas the baseline model showed incomplete detections. Under low-light conditions (Figure 8c), DASH-YOLO maintained clearer fitting

responses than YOLOv8x in the selected examples. In complex scenes (Figure 8d) involving structural clutter and diverse background interference, DASH-YOLO produced tighter bounding boxes and fewer missed detections, indicating improved robustness within the evaluated validation images.

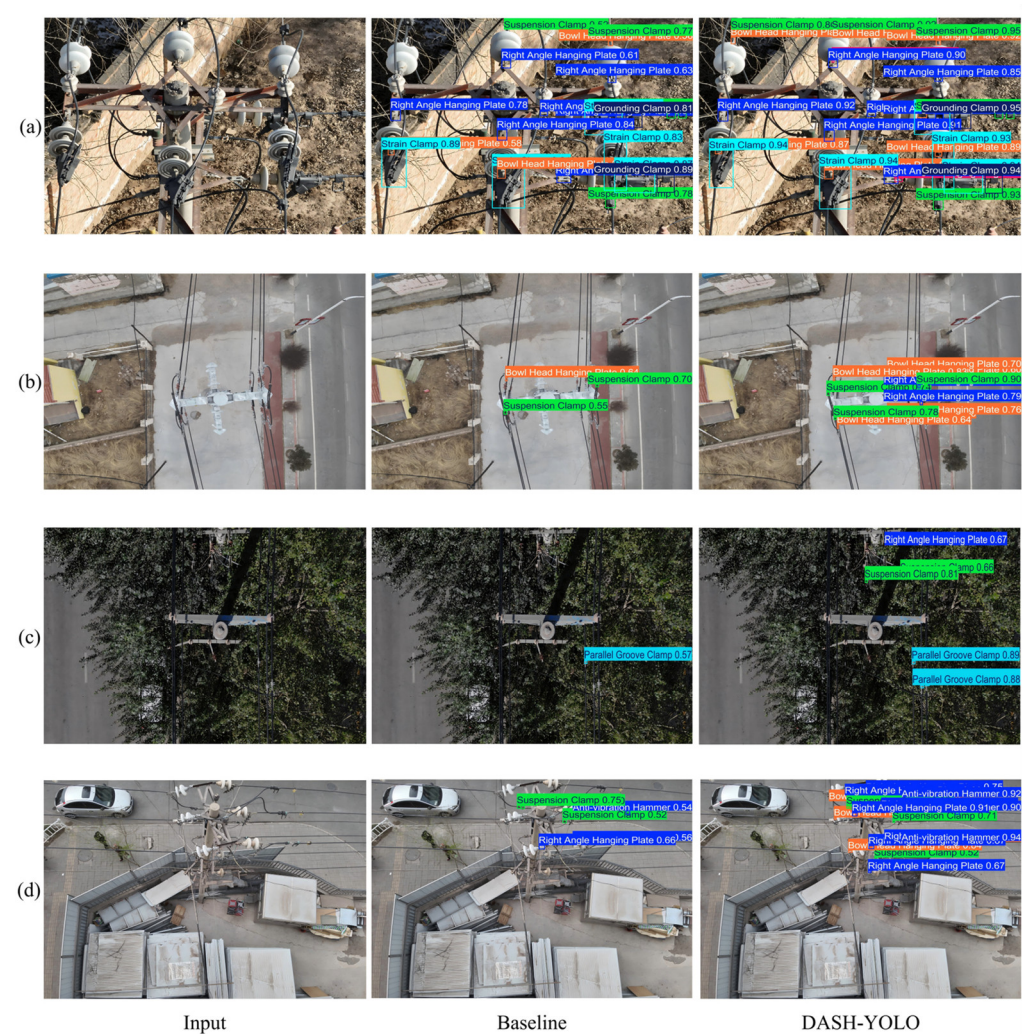


Figure 8. Qualitative comparison of detection performance between baseline YOLOv8x and DASH-YOLO under various scenarios: (a) dense target distribution; (b) sparse target distribution; (c) low-light conditions; and (d) complex background interference.

These qualitative results suggest that DASH-YOLO improves detection behavior over the YOLOv8x baseline in the illustrated validation examples, especially for small targets, adverse lighting, and cluttered backgrounds. Nevertheless, the visualization results are qualitative and dataset-specific; they should be interpreted together with the quantitative validation results and the stated limitations.

6. Conclusions

In this paper, DASH-YOLO is proposed as a deep learning framework for small fitting detection in transmission-line inspection images. By integrating a Dynamic Sparse Hyper Graph module for adaptive long-range dependency modeling and a Hierarchical Attention mechanism for progressive feature refinement, the proposed framework improves the detection of small-scale fittings in complex aerial imagery. A dedicated aerial inspection dataset covering ten typical fitting categories is constructed to support systematic evaluation. Experimental results show that DASH-YOLO achieves 90.37%

mAP@0.5, 74.74% mAP@0.5:0.95, 90.80% recall, and 85.49% precision, outperforming the YOLOv8x baseline and the evaluated comparison methods. Meanwhile, the model maintains practical efficiency with 95.73 GFLOPs, 65.73 MB parameters, and 62.99 FPS at the 640×640 input resolution. These results indicate that DASH-YOLO provides an effective accuracy-efficiency balance for UAV-assisted transmission-line fitting detection.

Although DASH-YOLO demonstrates favorable validation performance, several aspects can be further improved. The current DSHG module predicts a global density coefficient for each feature map, which controls the overall residual strength but does not yet provide spatially adaptive density modeling within the same image. In addition, the hypergraph reasoning module is mainly applied to the deepest P5 feature level, while extremely small fittings may benefit from additional lightweight relational modeling at shallower high-resolution levels. Future work will investigate spatially adaptive density prediction, multi-level lightweight hypergraph reasoning, temporal feature aggregation for UAV video inspection, deployment-oriented compression, multi-seed statistical testing, external small-object validation, and broader field evaluation under diverse weather, illumination, sensor, altitude, and geographic conditions.

Author Contributions: Conceptualization, Y.L. and Y.J.; methodology, Y.L., Y.J. and C.Z.; software, Y.J. and P.X.; validation, Y.J., P.X. and W.H.; formal analysis, Y.J. and C.Z.; investigation, Y.L. and W.H.; resources, Y.L.; data curation, P.X. and W.H.; writing—original draft preparation, Y.J.; writing—review and editing, Y.L., C.Z. and W.H.; visualization, Y.J. and P.X.; supervision, Y.L.; project administration, Y.L.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 62064008, and the Hubei Provincial Natural Science Foundation, grant number JCZRMS202600200.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yang, L.; Fan, J.F.; Liu, Y.H.; Li, E.; Peng, J.Z.; Liang, Z.Z. A Review on State-of-the-Art Power Line Inspection Techniques. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 9350–9365. [\[CrossRef\]](#)
2. Li, Y.; Dong, X.; Ding, Q.; Xiong, Y.; Liao, H.; Wang, T. Improved A-STAR Algorithm for Power Line Inspection UAV Path Planning. *Energies* **2024**, *17*, 5364. [\[CrossRef\]](#)
3. Tang, G.; Ni, J.; Zhao, Y.; Gu, Y.; Cao, W. A Survey of Object Detection for UAVs Based on Deep Learning. *Remote Sens.* **2024**, *16*, 149. [\[CrossRef\]](#)
4. Wang, Y.; Zou, H.; Yin, M.; Zhang, X. SMFF-YOLO: A Scale-Adaptive YOLO Algorithm with Multi-Level Feature Fusion for Object Detection in UAV Scenes. *Remote Sens.* **2023**, *15*, 4580. [\[CrossRef\]](#)
5. Min, L.; Fan, Z.; Lv, Q.; Reda, M.; Shen, L.; Wang, B. YOLO-DCTI: Small Object Detection in Remote Sensing Base on Contextual Transformer Enhancement. *Remote Sens.* **2023**, *15*, 3970. [\[CrossRef\]](#)
6. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017*; IEEE: New York, NY, USA, 2017; pp. 936–944. [\[CrossRef\]](#)
7. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: New York, NY, USA, 2018; pp. 8759–8768. [\[CrossRef\]](#)
8. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: Cham, Switzerland, 2018; Volume 11211, pp. 3–19. [\[CrossRef\]](#)
9. Wang, Q.; Wu, B.; Zhu, P.; Li, W.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020*; IEEE: New York, NY, USA, 2020; pp. 11531–11539. [\[CrossRef\]](#)

10. Liu, Y.; Shao, Z.; Hoffmann, N. Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions. *arXiv* **2021**, arXiv:2112.05561. [[CrossRef](#)]
11. Yang, X.; Xue, S.; Li, L.; Li, S.; Fang, Y.; Zhang, X.; Huang, X. PAMFPN: Position-Aware Multi-Kernel Feature Pyramid Network with Adaptive Sparse Attention for Robust Object Detection in Remote Sensing Imagery. *Remote Sens.* **2025**, *17*, 2213. [[CrossRef](#)]
12. Liu, G.; Zhang, S.; Wang, L.; Li, X.; Li, G. Research on Mechanical Automatic Food Packaging Defect Detection Model Based on Improved YOLOv5 Algorithm. *PLoS ONE* **2025**, *20*, e0321971. [[CrossRef](#)] [[PubMed](#)]
13. Mantach, S.; Lutfi, A.; Moradi Tavasani, H.; Ashraf, A.; El-Hag, A.; Kordi, B. Deep Learning in High Voltage Engineering: A Literature Review. *Energies* **2022**, *15*, 5005. [[CrossRef](#)]
14. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]
15. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017*; IEEE: New York, NY, USA, 2017; pp. 2980–2988. [[CrossRef](#)]
16. Chen, Y.; Deng, C.; Sun, Q.; Wu, Z.; Zou, L.; Zhang, G.; Li, W. Lightweight Detection Methods for Insulator Self-Explosion Defects. *Sensors* **2024**, *24*, 290. [[CrossRef](#)] [[PubMed](#)]
17. Redmon, J. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016*; IEEE: New York, NY, USA, 2016; pp. 779–788. [[CrossRef](#)]
18. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017*; pp. 7263–7271. [[CrossRef](#)]
19. Redmon, J. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767. [[CrossRef](#)]
20. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934. [[CrossRef](#)]
21. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976. [[CrossRef](#)]
22. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023*; IEEE: New York, NY, USA, 2023; pp. 7464–7475. [[CrossRef](#)]
23. Ahmed, M.D.F.; Mohanta, J.C. Autonomous Site Inspection of Power Transmission Line Insulators with Unmanned Aerial Vehicle System. *Electr. Power Compon. Syst.* **2024**, 1–24. [[CrossRef](#)]
24. Lu, Z.; Meng, X.; Li, J.; Yu, H.; Wang, S.; Qu, Z.; Wang, J. Detection of Photovoltaic Arrays in High-Spatial-Resolution Remote Sensing Images Using a Weight-Adaptive YOLO Model. *Energies* **2025**, *18*, 1916. [[CrossRef](#)]
25. Vieira-e-Silva, A.L.B.; de Castro Felix, H.; de Menezes Chaves, T.; Simões, F.P.M.; Teichrieb, V.; Santos, M.M.D. STN PLAD: A Dataset for Multi-Size Power Line Assets Detection in High-Resolution UAV Images. In *Proceedings of the 2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Gramado, Rio Grande do Sul, Brazil, 18–22 October 2021*; IEEE: New York, NY, USA, 2021; pp. 215–222. [[CrossRef](#)]
26. Chen, C.; Yuan, G.W.; Zhou, H.; Ma, Y. Improved YOLOv5s Model for Key Components Detection of Power Transmission Lines. *Math. Biosci. Eng.* **2023**, *20*, 7738–7760. [[CrossRef](#)] [[PubMed](#)]
27. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020*; IEEE: New York, NY, USA, 2020; pp. 10778–10787. [[CrossRef](#)]
28. Liang, J.; Wang, F.; Chen, J.; Huang, H.-Y.; Dou, Z.-F. Aerial Small Target Detection Algorithm Based on Cross-Scale Separated Attention. *PLoS ONE* **2025**, *20*, e0337318. [[CrossRef](#)] [[PubMed](#)]
29. Ouyang, D.; He, S.; Zhan, J.; Guo, H.; Huang, Z.; Luo, M.L.; Zhang, G.-L. Efficient Multi-Scale Attention Module with Cross-Spatial Learning. In *Proceedings of ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023*; IEEE: New York, NY, USA, 2023; pp. 1–5. [[CrossRef](#)]
30. Zhou, C.; Lu, Z.; Lv, Z.; Meng, M.; Tan, Y.; Xia, K.; Liu, K.; Zuo, H. Metal Surface Defect Detection Based on Improved YOLOv5. *Sci. Rep.* **2023**, *13*, 20803. [[CrossRef](#)] [[PubMed](#)]
31. Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; Liu, Z. Dynamic Convolution: Attention Over Convolution Kernels. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020*; IEEE: New York, NY, USA, 2020; pp. 11027–11036. [[CrossRef](#)]
32. Zhang, P.; Xu, H.; Tian, T.; Gao, P.; Tian, J. SFRE-Net: Scattering Feature Relation Enhancement Network for Aircraft Detection in SAR Images. *Remote Sens.* **2022**, *14*, 2076. [[CrossRef](#)]
33. Qu, S.; Dang, C.; Chen, W.; Liu, Y. SMA-YOLO: An Improved YOLOv8 Algorithm Based on Parameter-Free Attention Mechanism and Multi-Scale Feature Fusion for Small Object Detection in UAV Images. *Remote Sens.* **2025**, *17*, 2421. [[CrossRef](#)]
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: New York, NY, USA, 2018; pp. 7132–7141. [[CrossRef](#)]

35. Ilias, B.; Abdelaziz, B.; Anas, E.-N.; Douzi, S.; Douzi, H. Leveraging RegNet and CBAM for Precise Detection of Honey Adulteration Using Thermal Image Analysis. *Sci. Rep.* **2025**, *15*, 36555. [[CrossRef](#)] [[PubMed](#)]
36. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019*; IEEE: New York, NY, USA, 2019; pp. 1971–1980. [[CrossRef](#)]
37. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021*; IEEE: New York, NY, USA, 2021; pp. 13708–13717. [[CrossRef](#)]
38. Li, X. Ship Segmentation via Combined Attention Mechanism and Efficient Channel Attention High-Resolution Representation Network. *J. Mar. Sci. Eng.* **2024**, *12*, 1411. [[CrossRef](#)]
39. Yaseen, M.U.; Nasralla, M.M.; Aslam, F.; Ali, S.S.; Khattak, S.B.A. A Novel Approach Based on Multi-Level Bottleneck Attention Modules Using Self-Guided Dropblock for Person Re-Identification. *IEEE Access* **2022**, *10*, 123160–123176. [[CrossRef](#)]
40. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2016**, arXiv:1609.02907. [[CrossRef](#)]
41. Hao, X.; Li, J.; Guo, Y.; Jiang, T.; Yu, M. Hypergraph Neural Network for Skeleton-Based Action Recognition. *IEEE Trans. Image Process.* **2021**, *30*, 2263–2275. [[CrossRef](#)] [[PubMed](#)]
42. Yang, Y.; Liu, Z.; Yu, Z. SA-STGCN: A Spectral-Attentive Spatio-Temporal Graph Convolutional Network for Wind Power Forecasting with Wavelet-Enhanced Multi-Scale Learning. *Energies* **2025**, *18*, 5315. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.