information

MDPI

Article

GhostNetV2-YOLO: A Lightweight Detector for Multi-View Aesthetic Object Detection in Home Environments

Kaiwen Qiu [1], Yixuan Tu [2], Xin Zhou [3], Yiting Wang [4], Yiqun Tan [1,*] and Wenquan Huang [5]

[1] School of Modern Home Furnishings, Jiangxi Environmental Engineering Vocational College, Ganzhou 341002, China
[2] School of Fine Arts and Design, Rongjiang Campus, Gannan Normal University, Ganzhou 341000, China
[3] School of International Education, Krirk University, Bangkok 10220, Thailand
[4] School of Humanities and Law, Gannan University of Science and Technology, Ganzhou 341000, China
[5] School of Artificial Intelligence, Anhui University, Hefei 230601, China
* Correspondence: tyq15083570222@outlook.com

Abstract

With the accelerated progress of computational aesthetics and digital interior design, the demand for real-time and precise detection of aesthetic objects on edge devices has become increasingly pressing in applications such as intelligent design assistance, domestic aesthetic assessment, and augmented reality-based interior staging. As a core task in digital home aesthetics governance, virtual interior furnishing, household cultural archive development, and automated aesthetic evaluation, multi-view aesthetic object detection plays an essential role. However, this task still faces substantial difficulties arising from pronounced viewpoint variation, scale inconsistency, reflective materials, intricate decorative patterns, and cluttered indoor scenes. To address these issues, this study presents GhostNetV2-YOLO, a lightweight yet robust detection framework designed for accurate localization of aesthetic objects under unconstrained multi-view acquisition settings. The task is formally defined as closed-set detection of 10 pre-selected home aesthetic decorative items, including both planar decorative pieces and three-dimensional ornamental objects, and all performance claims are bounded within the horizontal bounding box detection paradigm. The framework incorporates three complementary components tailored to the target task. First, a task-adapted GhostNetV2 backbone is employed to enable efficient multi-scale feature extraction and long-range dependency modeling, with optimization specifically oriented toward structured aesthetic objects with stable global contours under viewpoint variation. Second, an improved Attention-based Intra-scale Feature Interaction (AIFI) module is introduced, integrating compressed QKV projection, linear attention, depthwise spatial refinement, and channel gating so that reflection-induced noise and background disturbance can be effectively reduced. Third, an enhanced Distance-IoU regression loss is adopted, in which explicit edge alignment and dynamic sample weighting are incorporated to improve boundary regression accuracy for rectangular and regularly contoured aesthetic objects. These designs jointly enhance contextual representation, boundary localization, and computational efficiency. Extensive experiments on two newly constructed multi-view aesthetic object datasets (AestheticHome-12K and AestheticHome-2K) demonstrate that the proposed detector achieves 94.80 $\pm$ 0.32%/94.20 $\pm$ 0.37% mAP@0.5, 96.30 $\pm$ 0.28%/95.60 $\pm$ 0.31% precision, and 94.70 $\pm$ 0.35%/93.80 $\pm$ 0.39% recall across two datasets (reported as mean $\pm$ standard deviation of 5 independent training runs with distinct random seeds), with only 2.89 M parameters and 6.0 GFLOPs. Statistical significance is verified via paired two-tailed t-tests with Bonferroni correction (adjusted $p < 0.05$) for all performance comparisons against baseline models. Compared with the YOLOv11n baseline, the method improves mAP@0.5 by 1.87–2.09 percentage points and recall by

Check for updates

Academic Editors: Ciprian Orhei, Muguras Mocofan and Radu Vasiu

Received: 19 July 2026
Revised: 9 August 2026
Accepted: 13 August 2026
Published: 14 August 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) license.

3.27–3.48 percentage points while reducing computational cost. Notably, it also achieves 79.2–80.5% mAP@0.5:0.95, outperforming the baseline by 4.7–4.9 percentage points, indicating significantly superior localization accuracy under stricter criteria. The proposed model achieves a remarkable balance between accuracy and efficiency, making it highly suitable for deployment on resource-constrained edge devices commonly used in digital design and home aesthetic monitoring systems. The results indicate that combining lightweight long-range feature extraction optimized for rigid aesthetic objects, compact attention-based feature interaction for interference suppression, and geometry-aware regression tailored for aesthetic targets provides an effective and efficient solution for robust aesthetic object detection in real-world computational aesthetics and digital interior design applications.

Keywords: aesthetic object detection; multi-view robustness; GhostNetV2-YOLO; GhostNetV2; attention mechanism; ImDIoU; computational aesthetics; digital interior design; edge computing

1. Introduction

In recent years, computational aesthetics has emerged as a vibrant interdisciplinary research area spanning computer vision, art design, cultural heritage, and intelligent human–computer interaction. In this setting, automatic aesthetic object detection serves as a fundamental component for home aesthetic digitization, virtual interior staging, automated design assistance, home cultural archive construction, and augmented reality art appreciation. Distinct from industrial products or ordinary daily objects, we formally define "home aesthetic objects" in this work as a closed set of 10 pre-selected decorative items commonly used for interior aesthetic display, including both planar decorative pieces (decorative plates, open books, polaroid cameras) and three-dimensional ornamental objects (ceramic vases, desktop telescopes, toy xylophones, spring ornaments, brain models, dinosaur skeleton models, toy guns). These items share stable global contour structures and serve dedicated aesthetic display functions in home environments, but they also present particular visual properties that make their detection considerably more difficult. Notably, not all categories are strictly planar or perfectly rectangular; therefore, the proposed edge-aligned regression loss is designed to strengthen boundary fitting for regularly contoured objects, and all conclusions in this work are constrained to horizontal bounding box-based detection scenarios.

Object detection has progressed rapidly alongside the development of deep learning. Earlier two-stage detectors, exemplified by Faster R-CNN, obtain strong localization performance through the combination of proposal generation and region-wise classification; however, their computational burden is often too high for real-time use [1]. By contrast, one-stage detectors represented by the YOLO series offer a more practical balance between detection accuracy and inference speed through the integration of localization and classification within a unified architecture [2–7]. To strengthen robustness across scales, feature pyramid architectures such as FPN and PANet improve the propagation and aggregation of multi-scale features [8,9]. More recently, transformer-based detectors, including DETR, Deformable DETR, and RT-DETR, have demonstrated the value of global context modeling and long-range dependency learning in object detection [10–12]. Meanwhile, lightweight backbones such as MobileNetV2, MnasNet, EfficientNet, GhostNet, and GhostNetV2 have been introduced to decrease parameter counts and floating-point operations for mobile and edge deployment scenarios [13–17]. Attention modules, including SE, CBAM, Coordinate Attention, and EMA, further enhance feature discrimination by highlighting informative

channels or spatial regions [18–21]. In addition, IoU-based losses, such as GIoU, DIoU, and Wise-IoU, have significantly improved the stability of regression and the convergence process [22–24]. Notwithstanding these developments, current methods still reveal several shortcomings in multi-view aesthetic object detection. Conventional lightweight detectors frequently compromise global contextual modeling under severe viewpoint variation. Standard attention modules may introduce non-negligible computational overhead or show limited ability in suppressing complicated indoor interference. Furthermore, commonly used box regression losses do not explicitly impose edge-level alignment, although such alignment is especially critical for rigid and planar aesthetic objects observed from oblique perspectives. In addition, most existing studies in aesthetic computing concentrate on image-level aesthetic assessment or large-scale art analysis, rather than on the fine-grained detection of small aesthetic objects within complex home environments [25–32].

To address these challenges, this paper presents GhostNetV2-YOLO, a lightweight and robust YOLO-based detector developed specifically for multi-view aesthetic object detection. The proposed method incorporates three complementary improvements that are tailored to the geometric and visual characteristics of aesthetic objects in diverse indoor scenes. First, the original backbone is substituted with a task-adapted GhostNetV2 [17], which integrates efficient ghost feature generation and decoupled fully convolutional attention to model the long-range structural dependencies of rigid planar aesthetic objects while maintaining low computational cost. Second, an improved Attention-based Intra-scale Feature Interaction (AIFI) module, derived from RT-DETR [12], is introduced to strengthen global–local feature interaction in a manner compatible with lightweight detection frameworks, with dedicated optimizations for suppressing reflections and background clutter. Third, an enhanced Distance-IoU loss is constructed through the incorporation of an explicit edge alignment term and a dynamic weighting mechanism, thereby improving boundary fitting and regression robustness for large, planar, and perspective-distorted aesthetic objects. Collectively, these components are intended to enhance detection accuracy, robustness, and deployment efficiency in multi-view aesthetic object scenarios.

More specifically, the proposed framework employs a three-scale detection pipeline built upon P3, P4, and P5 feature maps, together with semantic enhancement modules in the neck. The task-adapted GhostNetV2 backbone yields efficient multi-scale representations and reliable long-range feature extraction under viewpoint variation. The improved AIFI block compresses QKV channels, applies linear attention to avoid the quadratic complexity of standard self-attention, introduces depthwise convolution for spatial refinement, and employs channel gating for adaptive feature recalibration. This design preserves holistic structural cues while significantly reducing computational overhead. In the regression branch, the enhanced DIoU loss extends conventional center-distance-aware optimization by explicitly penalizing misalignment of the four box edges and by dynamically reweighting samples according to IoU statistics, which improves localization stability in reflective, cluttered, and oblique-view conditions.

The main contributions of this paper are summarized as follows:

- We propose GhostNetV2-YOLO, a lightweight YOLO-based detector specifically tailored for unconstrained multi-view aesthetic object detection in home environments, which integrates a task-adapted GhostNetV2 backbone optimized for contour-stable aesthetic object representation.
- We design an improved Attention-based Intra-scale Feature Interaction (AIFI) module with linear attention and interference-aware channel gating, which enhances global–local feature interaction while maintaining low computational complexity and effectively suppresses reflection and background clutter.

- We develop an enhanced Distance-IoU (ImDIoU) loss incorporating explicit edge alignment and dynamic sample weighting, which significantly improves the boundary localization accuracy of rectangular and irregular aesthetic objects under oblique views.
- We construct and release two new multi-view aesthetic object datasets (AestheticHome-12K and AestheticHome-2K) containing 14,000 annotated images captured in diverse real home environments, covering 10 common aesthetic object categories with viewpoint annotations from 0° to 60°.
- We provide full reproducibility support including fixed dataset split protocols, exact code version archives, detailed training configuration files, and complete statistical validation based on 5 independent training runs with significance testing.
- Extensive experiments with rigorous statistical verification on the proposed datasets demonstrate that the proposed method outperforms state-of-the-art lightweight detectors in both accuracy and efficiency, providing a practical solution for edge-enabled computational aesthetics and digital interior design systems.

Experiments are conducted on the two newly constructed multi-view aesthetic object datasets, AestheticHome-12K and AestheticHome-2K, using a unified training setting under PyTorch 2.2.2 and Python 3.10 on an NVIDIA RTX 4090 platform. The proposed model achieves 94.80% mAP@0.5, 96.30% precision, and 94.70% recall on AestheticHome-12K, and 94.20% mAP@0.5, 95.60% precision, and 93.80% recall on AestheticHome-2K, while maintaining only 2.89 M parameters and 6.0 GFLOPs. Relative to the YOLOv11n baseline, the method improves mAP@0.5 by 1.87–2.09 percentage points and recall by 3.27–3.48 percentage points, while slightly reducing computational complexity. Ablation and robustness analyses further confirm that the task-adapted GhostNetV2, the improved AIFI module, and the enhanced DIoU loss each contribute meaningfully to performance gains, and that their combination yields strong robustness across different viewpoints. These results demonstrate that the proposed framework provides an effective and efficient solution for real-time multi-view aesthetic object detection in computational aesthetics and digital interior design systems.

2. Related Work

Object detection in complex visual scenes is generally approached through three principal technical directions: efficient detector design, discriminative feature modeling, and accurate bounding-box regression. In the context of multi-view aesthetic object detection, these directions become particularly critical, because the task must concurrently accommodate severe viewpoint variation, scale changes, complex indoor background clutter, reflective interference, and stringent edge deployment requirements. Although existing research in general object detection and computational aesthetics offers valuable foundations, the corresponding methods do not fully satisfy the combined demands for robustness and efficiency in unconstrained real home environments. Accordingly, this section reviews previous studies from the standpoint of the core issues addressed in this work, including real-time multi-scale detection, lightweight feature extraction and attention modeling, regression optimization for precise localization, and computational aesthetics and aesthetic object analysis.

2.1. Real-Time Object Detection and Multi-Scale Feature Fusion

A widely adopted strategy in object detection is to enhance localization and classification performance by reinforcing multi-scale representation and feature aggregation. Ren et al. introduced Faster R-CNN, a representative two-stage detector that incorporated the Region Proposal Network and substantially improved detection accuracy by unifying proposal generation and object recognition within a single framework [1]. This approach

overcame the inefficiency associated with earlier hand-crafted region proposal pipelines, although its multi-stage inference procedure constrained its suitability for real-time deployment. In response to the demand for higher speed, Redmon et al. proposed YOLO, which recast object detection as a single regression task and achieved real-time performance through end-to-end prediction [2]. On this foundation, Redmon and Farhadi subsequently developed YOLO9000 and YOLOv3, further enhancing category scalability, anchor-based prediction, and multi-scale detection capability [3,4]. Collectively, these studies established the practical strengths of one-stage detectors in achieving a favorable balance between speed and accuracy.

To further improve robustness across scales, Lin et al. proposed the Feature Pyramid Network (FPN), which integrated low-level spatial details with high-level semantic information through a top-down pathway and lateral connections [8]. This architecture effectively mitigated the problem of insufficient semantic representation in shallow layers and improved the detection of small objects. Building upon FPN, Liu et al. presented PANet, which introduced bottom-up path augmentation to shorten the information path between shallow and deep layers, thereby enhancing localization performance and feature propagation [9]. Lin et al. also developed RetinaNet with focal loss, addressing the pronounced imbalance between foreground and background samples in dense detection and improving the training stability of one-stage detectors [33]. More recently, Wang et al. proposed YOLOv9 with programmable gradient information, which improved optimization flexibility and the effectiveness of feature learning [6], while another group led by Wang introduced YOLOv10 as an end-to-end real-time detector that further simplified the prediction pipeline and increased efficiency [7]. Although these methods have advanced general-purpose object detection, they still encounter notable difficulties under multi-view aesthetic object conditions, where extreme oblique viewpoints, reflective surfaces, and substantial indoor background interference remain challenging.

Global context modeling has also become increasingly important for robust detection. Carion et al. proposed DETR, which introduced transformer-based set prediction into object detection and eliminated hand-crafted anchor design and non-maximum suppression dependency through end-to-end bipartite matching [10]. This work solved the limitation of local receptive fields in convolutional detectors, but the convergence speed and computational cost remained challenging. Zhu et al. proposed Deformable DETR, which improved DETR by using sparse deformable attention to reduce computational burden and accelerate training [11]. More recently, Zhao et al. proposed RT-DETR, which combined efficient hybrid encoding with real-time detection and demonstrated that transformer-based global modeling can be practical in speed-sensitive scenarios [12]. RT-DETR is particularly relevant to our work because it inspired the use of AIFI-style feature interaction for efficient context aggregation. However, even these improved transformer-based approaches remain relatively expensive for lightweight deployment in digital design systems. Motivated by these studies, our work adopts the YOLO framework as the basic detector and strengthens its multi-view robustness through task-adapted GhostNetV2 backbone replacement, efficient intra-scale attention interaction optimized for interference suppression, and improved regression design tailored for aesthetic objects.

2.2. Lightweight Backbones and Attention Mechanisms

A major challenge in aesthetic object detection is how to maintain strong feature representation under constrained computational budgets. Conventional high-capacity backbones often improve accuracy but are difficult to deploy in edge devices used in digital design environments, such as embedded monitoring systems, mobile design platforms, or smartphone-based aesthetic analysis tools. To address this issue, Sandler et al.

proposed MobileNetV2, which introduced inverted residuals and linear bottlenecks to reduce parameters and computation while preserving representational power [13]. This work solved the inefficiency of standard convolutional networks in mobile scenarios. Tan et al. then proposed MnasNet, which used platform-aware neural architecture search to optimize the trade-off between latency and accuracy [14]. Compared with hand-designed lightweight networks, MnasNet provided a more deployment-oriented solution. Tan and Le further developed EfficientNet, which introduced compound scaling of depth, width, and resolution, thereby improving accuracy–efficiency balance in a principled manner [15]. These lightweight networks advanced mobile vision, but they primarily focused on general image recognition and did not explicitly model long-range dependencies, which are critical under large viewpoint changes for rigid planar aesthetic objects.

To generate efficient features more directly, Han et al. proposed GhostNet, arguing that many conventional feature maps contain redundancy and can be approximated by cheap linear operations [16]. This design greatly reduced model complexity while maintaining competitive performance. Building on this idea, Tang et al. proposed GhostNetV2, which incorporated Decoupled Fully Convolutional (DFC) attention into the Ghost framework to enhance long-range dependency modeling with low overhead [17]. Compared with GhostNet, GhostNetV2 addressed the limitation of insufficient global contextual modeling in lightweight backbones, making it especially suitable for scenes where object structure must remain stable across different viewpoints. Because aesthetic objects are typically rigid and planar, and their global contours remain informative even under local appearance changes or partial occlusion, GhostNetV2 provides an effective basis for robust feature extraction in our task. In this work, we further adapt GhostNetV2 for detection tasks by adjusting the output stride configuration and optimizing the attention block distribution to better capture multi-scale structural features of aesthetic objects.

Attention mechanisms have also been widely used to improve feature discrimination. Hu et al. proposed the Squeeze-and-Excitation (SE) block, which adaptively recalibrates channel responses through global pooling and gating, thereby improving channel selectivity with minimal overhead [18]. Woo et al. proposed CBAM, which extended this idea by sequentially combining channel attention and spatial attention, improving feature refinement in both dimensions [19]. Hou et al. developed Coordinate Attention, which embedded long-range positional information into channel attention, making it better suited to mobile networks and dense prediction tasks [20]. Ouyang et al. further presented an efficient multi-scale attention module with cross-spatial learning, which improved feature interaction across different spatial regions while maintaining computational efficiency [21]. These studies demonstrated that attention mechanisms can effectively suppress irrelevant background responses and emphasize informative object regions. Nevertheless, standard self-attention is hindered by quadratic complexity, and many plug-in attention modules are not optimized for the strict efficiency demands of real-time detection. Moreover, they are not specifically designed to mitigate the reflection and background interference frequently encountered in home environments.

For this reason, compressed and linearized attention mechanisms have attracted increasing interest. Vaswani et al. first demonstrated the effectiveness of self-attention in capturing long-range dependencies for sequence modeling [34], yet the quadratic complexity of vanilla attention restricts its direct application in lightweight visual detectors. RT-DETR partially addressed this limitation through efficient hybrid encoding and attention-based feature interaction [12]. Motivated by these advances, our work develops an improved AIFI module for intra-scale feature interaction. In comparison with conventional attention blocks, the proposed module compresses QKV channels, employs linear attention to lower computational complexity, incorporates depthwise convolution for spatial refine-

ment, and adopts interference-aware channel gating for adaptive modulation. This design directly addresses the issue highlighted in previous studies, namely how to retain global context modeling while ensuring that the detector remains sufficiently lightweight for resource-constrained multi-view aesthetic object applications.

2.3. IoU-Based Bounding Box Regression Losses

Another critical challenge in aesthetic object detection is achieving accurate localization under oblique viewing angles, across large planar surfaces, and in the presence of boundary interference from surrounding furniture or reflective materials. In many situations, a predicted box may overlap with the target while still failing to align accurately with the true borders of the aesthetic object, thereby reducing practical utility in applications such as aesthetic analysis, digital design, and augmented reality systems. To overcome the limitations of conventional regression objectives, Rezatofighi et al. proposed Generalized IoU (GIoU), which extended standard IoU by introducing a penalty term derived from the smallest enclosing box [22]. This approach addressed the gradient vanishing issue when the predicted and ground-truth boxes do not overlap. However, GIoU does not explicitly model the distance between box centers. Zheng et al. subsequently proposed DIoU loss, which incorporated a normalized center-distance penalty to accelerate convergence and improve localization efficiency [23]. Relative to GIoU, DIoU offered stronger geometric guidance during training and consequently became a widely adopted regression objective in real-time detectors.

Although DIoU improves center alignment, it still does not clearly distinguish between center-level proximity and precise edge-level fitting. For rigid planar objects such as aesthetic objects, accurate boundary alignment is especially important, because perspective distortion and reflective interference can easily lead to prediction offsets along one or more edges. To enhance sample weighting during regression, Tong et al. proposed Wise-IoU, which introduced a dynamic focusing mechanism that adaptively adjusts the contribution of different samples according to training status [24]. This method mitigated the adverse influence of low-quality or noisy samples and improved regression robustness. Nevertheless, Wise-IoU does not explicitly penalize edge misalignment, which remains a crucial issue for aesthetic object localization. Therefore, previous regression losses have each addressed only part of the problem: GIoU resolves non-overlap, DIoU enhances geometric convergence, and Wise-IoU introduces adaptive weighting. Building upon these insights, our work proposes an enhanced DIoU loss that integrates center-distance-aware optimization with explicit edge alignment and dynamic sample weighting, with the goal of achieving more stable and precise localization for large, planar, and perspective-distorted aesthetic objects.

2.4. Computational Aesthetics and Aesthetic Object Analysis

Within the broader domain of computational aesthetics, most existing research has concentrated on image-level aesthetic assessment or large-scale art analysis, rather than the fine-grained detection of small aesthetic objects. Datta et al. introduced a computational method for evaluating photographic aesthetic quality, demonstrating that machine learning models can be trained to predict human aesthetic preferences [25]. Marchesotti et al. further developed a framework for image aesthetic evaluation based on high-level visual descriptors [26]. More recently, deep learning techniques have been widely applied to aesthetic assessment and have achieved state-of-the-art results on multiple benchmarks [27–29]. However, these studies mainly address global image aesthetics, rather than the detection and analysis of individual aesthetic objects embedded in complex scenes.

A number of studies have more directly examined aesthetic object analysis in art and design contexts. Saleh and Elgammal investigated large-scale classification of fine-

art paintings, showing that deep learning can effectively recognize diverse artistic styles and genres [30]. Kim et al. introduced a dataset for detecting visual sentiment from art images [31], while Yi et al. proposed a method for artistic image aesthetics assessment [32]. Nevertheless, these approaches are tailored to high-quality art images captured in relatively controlled settings and do not address the challenges associated with multi-view detection in cluttered home environments. This limitation suggests that existing computational aesthetics methods cannot be directly transferred to the practical task of home aesthetic object detection considered in this work. Therefore, although our study is related to computational aesthetics, it is more precisely situated at the intersection of real-time object detection, lightweight model design, and digital interior design applications.

In recent years, edge computing has emerged as a key enabling technology for intelligent design systems, supporting real-time data processing and analysis directly at the acquisition site [35]. However, most existing edge-based object detectors are designed for general scenarios and do not consider the specific challenges of multi-view aesthetic object detection, such as extreme perspective distortion and reflective interference. This gap motivates our work to develop a lightweight and robust detector specifically optimized for digital design edge environments.

2.5. Summary and Motivation of the Present Study

The above studies provide important inspiration but also reveal clear limitations for multi-view aesthetic object detection. In terms of detector design, Faster R-CNN [1], YOLO [2–4], FPN [8], PANet [9], RetinaNet [33], YOLOv9 [6], YOLOv10 [7], DETR [10], Deformable DETR [11], and RT-DETR [12] have progressively improved speed, scale robustness, and global modeling, yet they do not fully resolve the combined challenges of extreme viewpoint variation, reflective interference, and edge deployment in unconstrained home environments. In terms of efficient representation learning, MobileNetV2 [13], MnasNet [14], EfficientNet [15], GhostNet [16], and GhostNetV2 [17] have reduced computational burden, while SE [18], CBAM [19], Coordinate Attention [20], and related modules [21] improved feature discrimination. However, many existing solutions either lack sufficient long-range dependency modeling optimized for rigid planar aesthetic objects or remain suboptimal in efficiency–robustness balance for this task. In terms of localization, GIoU [22], DIoU [23], and Wise-IoU [24] have improved regression quality, but none specifically target the edge alignment requirement of rectangular aesthetic objects under oblique viewpoints. Furthermore, there is a critical lack of publicly available multi-view datasets specifically designed for home aesthetic object detection in real environments.

Motivated by these observations, we propose GhostNetV2-YOLO, a lightweight and robust YOLO-based detector that combines a task-adapted GhostNetV2 backbone for efficient long-range feature extraction optimized for rigid aesthetic objects, an improved AIFI module for compact global–local context interaction with interference suppression, and an enhanced DIoU loss for geometry-aware and edge-sensitive localization tailored for aesthetic objects. We also contribute two new multi-view aesthetic object datasets to facilitate future research in this field. In this way, the proposed framework is directly inspired by the strengths of prior work while addressing their limitations in the specific context of multi-view aesthetic object detection.

3. Methodology

3.1. Overall Architecture

To improve multi-view aesthetic object detection under extreme viewpoint variation, scale change, reflection interference, and complex indoor background clutter, we propose GhostNetV2-YOLO, a lightweight detection framework based on YOLOv11n. YOLOv11n

is chosen as the baseline due to its excellent speed-accuracy trade-off for edge deployment. We retain its three-scale detection pipeline but make three key task-specific modifications: replacing the original CSPDarknet backbone with a task-adapted GhostNetV2, inserting the improved AIFI module after neck fusion for each scale, and replacing the standard CIoU loss with the proposed enhanced ImDIoU loss. As shown in Figure 1, the proposed method consists of three key components:

- A task-adapted GhostNetV2 backbone for efficient hierarchical feature extraction optimized for rigid planar aesthetic objects;
- An improved AIFI module for intra-scale feature refinement with interference suppression;
- An enhanced DIoU loss for more accurate bounding-box regression of aesthetic objects.

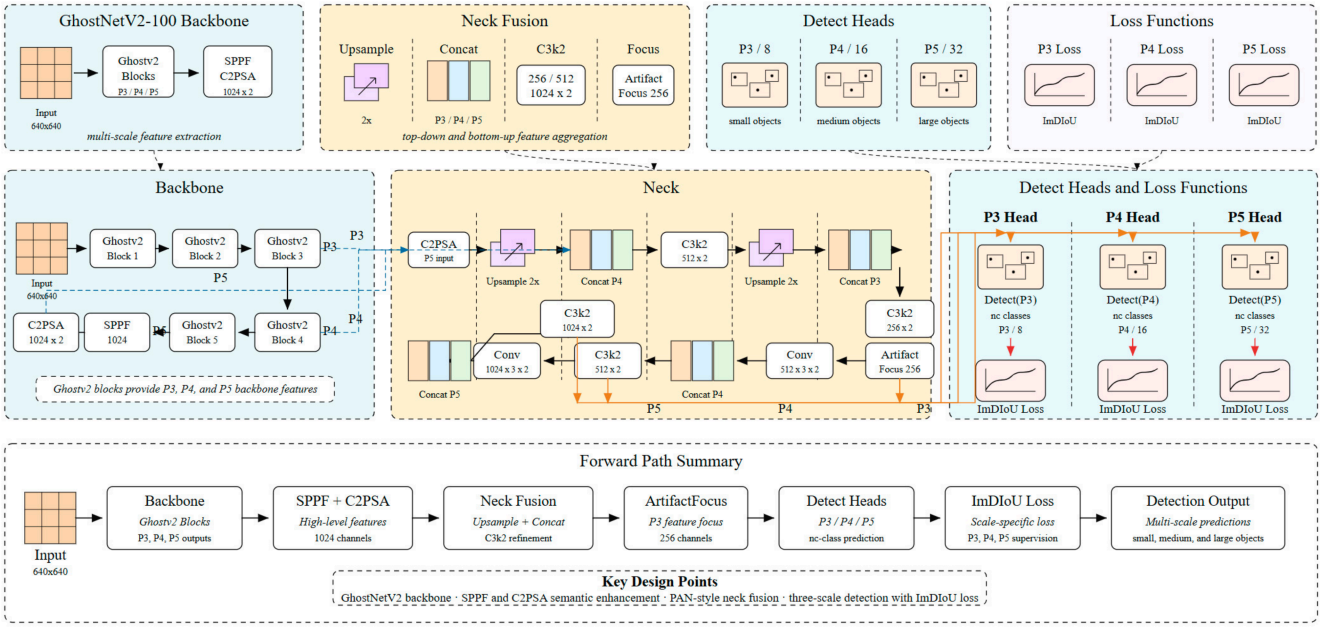


Figure 1. GhostNetV2-YOLO Architecture Roadmap. The framework consists of four main parts: task-adapted GhostNetV2-100 backbone for multi-scale feature extraction, PAN-style neck for top-down and bottom-up feature aggregation, improved AIFI modules for intra-scale feature refinement, three-scale detection heads for predicting small, medium, and large aesthetic objects, and ImDIoU loss for geometry-aware bounding box regression. Input resolution is 640×640, and output feature maps have strides of 8, 16, and 32 respectively.

Given an input image $I \in \mathbb{R}^{3\times640\times640}$, the network first extracts multi-scale feature maps through the task-adapted GhostNetV2 backbone. Three output feature levels, $P3(80 \times 80), P4(40 \times 40)$, and $P5(20 \times 20)$, are then delivered to the neck for top-down and bottom-up fusion. After feature aggregation, the improved AIFI attention refinement module is introduced before prediction to strengthen discriminative representation. Finally, three detection heads perform classification and localization at different scales.

The overall inference process can be written as

$$\{Y_3, Y_4, Y_5\} = \mathcal{H}(\mathcal{N}(\mathcal{B}(I))) \tag{1}$$

where $B(\cdot)$ denotes the task-adapted backbone, $N(\cdot)$ denotes the neck and feature refinement stage, and $H(\cdot)$ denotes the multi-scale detection heads.

Compared with the baseline detector, the proposed method improves both efficiency and accuracy. The task-adapted GhostNetV2 reduces redundant computation in feature extraction while enhancing long-range structural representation, the improved AIFI module

enhances global–local interaction with low complexity and suppresses irrelevant background and reflection interference, and the enhanced DIoU loss strengthens geometry-aware localization, especially for large planar aesthetic object targets.

3.2. Task-Adapted GhostNetV2 Backbone

3.2.1. Design Motivation

Backbone design is critical for aesthetic object detection because aesthetic object instances often exhibit rigid geometry, large spatial extent, and significant viewpoint distortion. A practical detector should therefore be capable of capturing both fine-grained local visual details and long-range structural cues while maintaining high computational efficiency. To achieve this goal, we replace the original backbone with a task-adapted GhostNetV2, which combines lightweight ghost feature generation with attention-enhanced spatial modeling tailored to rigid planar aesthetic objects.

Figure 2 presents the detailed architecture of the task-adapted GhostNetV2 and its integration into the YOLO detection backbone. The backbone is constructed by stacking GhostBottleneckV2 blocks, which operate in two modes: a standard Ghost mode in the early layers to maximize efficiency, and an attention mode in the deeper layers to strengthen long-range dependency modeling. Compared with the original classification-oriented GhostNetV2, we make two key adaptations for detection tasks: (1) we adjust the output stride configuration to produce feature maps at strides 8, 16, and 32 corresponding to P3, P4, and P5 levels; (2) we increase the number of attention-enhanced blocks in stages 4 and 5 to better capture global structural features of large planar aesthetic objects.

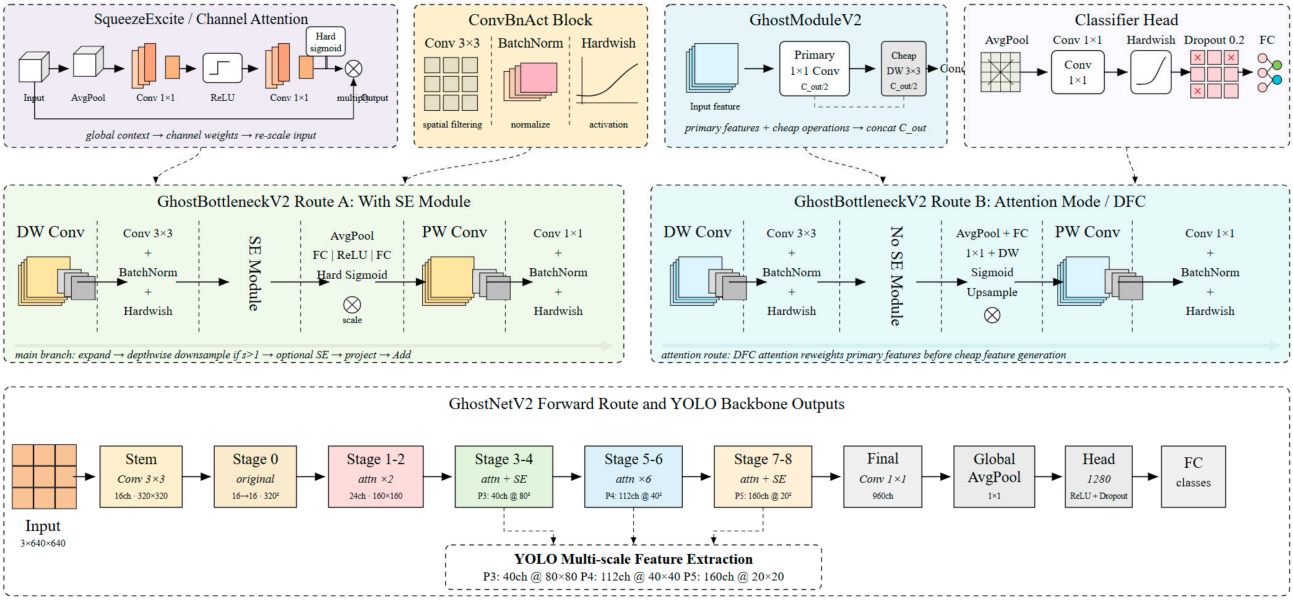


Figure 2. Task-Adapted GhostNetV2 Roadmap for GhostNetV2-YOLO Detection Backbone. The figure shows the core components of the task-adapted GhostNetV2, including GhostModuleV2, GhostBottleneckV2 with SE module (Route A) and DFC attention mode (Route B), and the forward path for multi-scale feature extraction. Stages 1–3 use standard Ghost blocks, while Stages 4–5 use an increased number of attention-enhanced blocks to improve long-range structural modeling of aesthetic objects.

3.2.2. Ghost Feature Generation

The key idea of GhostNetV2 is to reduce feature redundancy. Instead of generating all output channels using expensive standard convolutions, it first produces a small number of intrinsic feature maps and then generates additional feature maps through cheap operations.

For an input feature map X, the primary branch first computes

$$\mathrm{F}_p = f_p(\mathrm{X}) \tag{2}$$

where $f_p(\cdot)$ denotes a standard convolution, typically a 1×1 convolution. Then, a series of cheap linear operations are applied to obtain additional ghost features:

$$\mathrm{F}_g^{(i)} = \phi_i(\mathrm{F}_p),\ i = 1, \ldots, s-1 \tag{3}$$

where $\phi_i(\cdot)$ is usually a depthwise convolution. The final output is formed by channel concatenation:

$$\mathrm{Y} = Concat\left[\mathrm{F}_p, \mathrm{F}_g^{(1)}, \ldots, \mathrm{F}_g^{(s-1)}\right] \tag{4}$$

This mechanism greatly reduces parameter count and FLOPs while preserving feature diversity.

3.2.3. Attention-Enhanced GhostModuleV2

To improve long-range dependency modeling, GhostNetV2 introduces an attention mode based on decoupled fully connected (DFC) attention. Given an input $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, the module first downsamples the feature map:

$$\mathrm{X}_d = AvgPool(\mathrm{X}) \tag{5}$$

then applies a lightweight convolutional gating branch:

$$\mathrm{A} = \sigma(f_2(f_1(\mathrm{X}_d))) \tag{6}$$

where f_1 and f_2 include 1×1, (1,5) depthwise, and (5,1) depthwise convolutions, and $\sigma(\cdot)$ denotes the sigmoid function. After upsampling, the generated attention map modulates the primary features:

$$F_p = F_p \odot Up(A) \tag{7}$$

where $\odot$ denotes element-wise multiplication.

This design enables the backbone to enhance informative spatial regions while preserving the lightweight nature of ghost feature generation.

3.2.4. GhostBottleneckV2 for Hierarchical Representation

The backbone is built by stacking GhostBottleneckV2 blocks. Each bottleneck follows an inverted residual structure:

$$Y = G_{proj}(S(D(G_{exp}(X)))) + \mathcal{R}(X) \tag{8}$$

where G_{exp} and G_{proj} denote the expansion and projection Ghost modules, D denotes optional depthwise downsampling, S denotes an optional SE block, and $\mathcal{R}(X)$ denotes the identity shortcut branch for residual connection.

In our implementation, we use task-adapted GhostNetV2-100 as the backbone, which outputs three multi-scale feature maps: P3 (80×80), P4 (40×40), and P5 (20×20) for an input image of size 640×640. The first three stages use standard GhostBottleneckV2 blocks without attention to maximize efficiency, while the last two stages use an increased number of attention-enhanced blocks to capture long-range dependencies critical for viewpoint-invariant recognition of rigid planar aesthetic objects.

3.3. Neck and Multi-Scale Detection Pipeline

To fully exploit the hierarchical outputs of the backbone, we construct a multi-scale neck for semantic aggregation and feature interaction. The deepest feature map P_5 is first enhanced by an SPPF block and a C2PSA (Contextual Pyramid Spatial Attention) module from YOLOv11 [7], which aggregates multi-scale contextual information to improve large-object detection. C2PSA is a lightweight spatial attention module that applies parallel dilated convolutions with different dilation rates (1, 3, 5) to capture multi-scale contextual information, followed by channel-wise concatenation and a 1×1 convolution for feature fusion. This design is motivated by prior studies on atrous convolution for multi-scale context aggregation [36]. Then, a top-down pathway propagates high-level semantics to shallower layers through upsampling and concatenation:

$$P_4 = \mathcal{F}_4(Concat(Up(P_5), P_4)) \tag{9}$$

$$P_3 = \mathcal{F}_3(Concat(Up(P_4), P_3)) \tag{10}$$

where $\mathcal{F}_3$ and $\mathcal{F}_4$ denote convolutional fusion blocks consisting of C3k2 modules.

The improved AIFI module is inserted after the neck fusion stage for each scale (P3, P4, P5). This placement allows the module to refine the already fused high-level features before they are fed into the detection heads, achieving better performance with lower computational overhead compared to inserting attention in the backbone. The prediction stage outputs three scale-specific detection tensors:

$$Y_s = \mathcal{H}_s(P_3), \;\; Y_m = \mathcal{H}_m(P_4), \;\; Y_l = \mathcal{H}_l(P5) \tag{11}$$

where Y_s, Y_m, and Y_l correspond to small-scale, medium-scale, and large-scale targets, respectively.

3.4. Improved AIFI Module with Interference-Aware Channel Gating

3.4.1. Motivation

Although convolution-based fusion is effective for local feature aggregation, it is relatively limited in its ability to model long-range dependencies within the same feature scale. This shortcoming is particularly pronounced in multi-view aesthetic object detection, where large planar objects require both precise local boundary perception and global structural reasoning. Moreover, home environments often include substantial interference from glass reflections and cluttered indoor backgrounds, which may divert the detector's attention away from the true aesthetic targets. To address these challenges, we introduce an improved Attention-based Intra-scale Feature Interaction (AIFI) module that integrates compressed QKV projection, linear attention, interference-aware channel gating, and a lightweight feed-forward network.

Figure 3 illustrates the detailed architecture of the improved AIFI module, depicting the complete forward process from the input feature map to the output response.

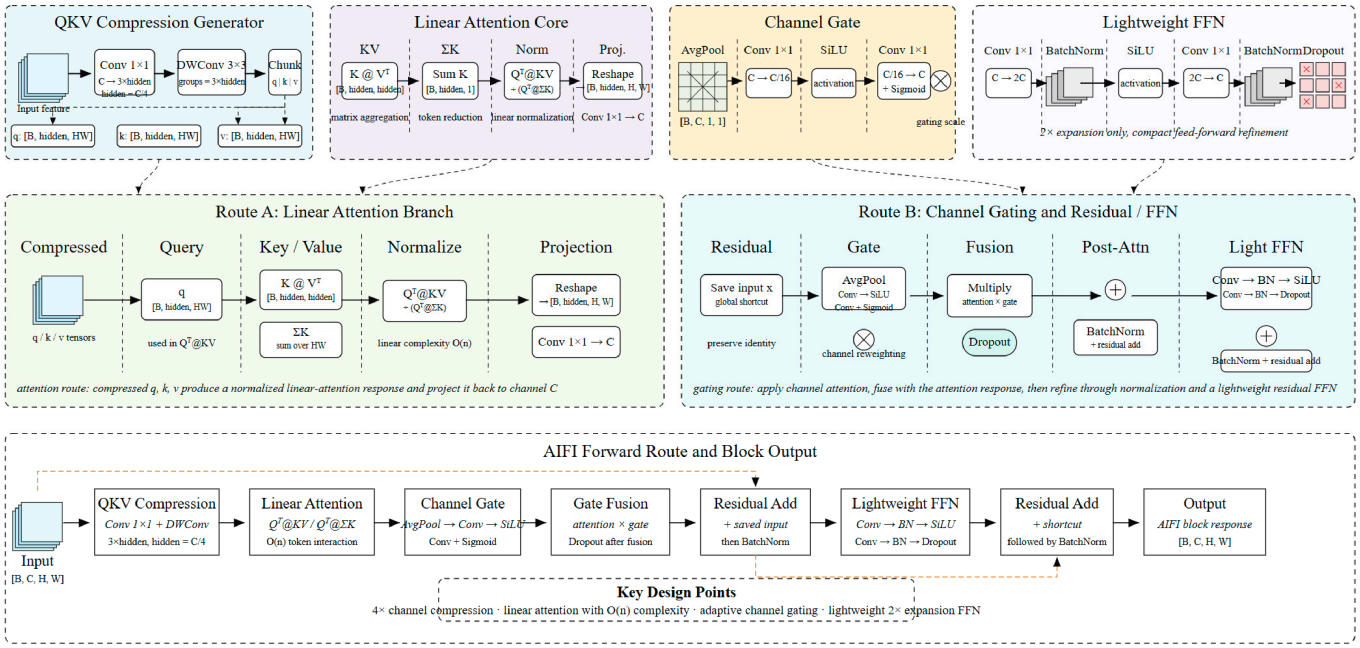


Figure 3. Improved AIFI Attention Block with Interference-Aware Channel Gating. The module consists of four core components: QKV compression generator, linear attention core with O(N) complexity, interference-aware adaptive channel gate, and lightweight expansion FFN.

3.4.2. Compressed QKV Projection with Local Refinement

Let the input feature map be $X \in \mathbb{R}^{B\times C\times H\times W}$. To reduce computational cost, a 1×1 convolution first projects the input into compressed query, key, and value tensors:

$$Z = Conv_{1\times1}(X) \in \mathbb{R}^{B\times 3d\times H\times W} \tag{12}$$

where

$$d = \frac{C}{4} \tag{13}$$

We choose a $4\times$ channel compression ratio based on ablation experiments, which balances computational efficiency and feature representation capacity (see Section 4.5.2 for detailed analysis).

To enrich local spatial information before attention computation, a depthwise 3×3 convolution is further applied:

$$\widetilde{Z} = \mathrm{DWConv}_{3\times3}(Z) \tag{14}$$

The resulting tensor is evenly split into three parts:

$$[Q, K, V] = Chunk(\widetilde{Z}) \tag{15}$$

and each part is reshaped into

$$Q, K, V \in \mathbb{R}^{B\times d\times N},\ N = HW \tag{16}$$

Following the linear attention formulation, a ReLU kernel is applied to ensure non-negativity:

$$Q \leftarrow ReLU(Q),\ K \leftarrow ReLU(K),\ V \leftarrow ReLU(V) \tag{17}$$

3.4.3. Linear Attention Computation

Unlike standard self-attention, which requires quadratic complexity $O(N^2)$, we adopt a linear attention mechanism to improve efficiency. First, key-value aggregation is computed as

$$\mathrm{M} = \mathrm{K}\mathrm{V}^\top \in \mathbb{R}^{B\times d\times d} \tag{18}$$

and the key summation term is

$$\mathrm{s} = \sum_{n=1}^{N} \mathrm{K}_{:,n} \in \mathbb{R}^{B\times d\times 1} \tag{19}$$

The attention output is then obtained by

$$\mathrm{O} = \frac{\mathrm{Q}^\top \mathrm{M}}{\mathrm{Q}^\top \mathrm{s} + \varepsilon} \tag{20}$$

where $\varepsilon = 10^{-6}$ is a small constant to avoid numerical instability. After reshaping, the attended feature map is

$$\mathrm{O} \in \mathbb{R}^{B\times d\times H\times W} \tag{21}$$

This formulation reduces the computational complexity from $O(N^2)$ to $O(N)$, making it suitable for high-resolution detection features.

3.4.4. Interference-Aware Channel Gating Mechanism

To improve channel-wise feature selectivity and suppress interference from reflections and background clutter, we introduce an interference-aware lightweight gating branch. Global context is first extracted by adaptive average pooling:

$$\mathrm{g} = GAP(\mathrm{X}) \in \mathbb{R}^{B\times C\times 1\times 1} \tag{22}$$

Then, two 1×1 convolutions generate channel weights:

$$\mathrm{w} = \sigma(W_2\delta(W_1\mathrm{g})) \tag{23}$$

where W_1: $C\to C/16$, W_2: $C/16\to C$, $\delta(\cdot)$ denotes the SiLU activation, and $\sigma(\cdot)$ is the sigmoid function. The gating mechanism learns to suppress channels that respond strongly to background clutter and reflections while enhancing channels that carry aesthetic object-specific structural information.

The attended features are projected back to the original channel dimension:

$$\mathrm{O}' = Conv_{1\times 1}(\mathrm{O}) \tag{24}$$

and recalibrated as

$$\widetilde{\mathrm{O}} = \mathrm{O}' \odot \mathrm{w} \tag{25}$$

3.4.5. Lightweight Feed-Forward Network

A residual connection is first applied after attention refinement:

$$\mathrm{X}_1 = \mathrm{BN}(\widetilde{\mathrm{O}}) + \mathrm{X} \tag{26}$$

Then, a lightweight feed-forward network with only $2\times$ expansion is employed to avoid redundant computation:

$$\mathrm{F} = W_4\delta(\mathrm{BN}(W_3\mathrm{X}_1)) \tag{27}$$

where W_3: $C \rightarrow 2C$ and W_4: $2C \rightarrow C$. The final output is given by

$$Y = BN(F) + X_1 \tag{28}$$

Compared with conventional transformer-style FFNs with 4× expansion, the proposed design reduces redundant computation while maintaining effective feature transformation.

3.4.6. Summary of the Improved AIFI Module

The improved AIFI module can be summarized as

$$Y = \mathcal{A}_{\mathrm{ImAIFI}}(X) \tag{29}$$

where $\mathcal{A}_{\mathrm{ImAIFI}}$ includes compressed QKV generation, depthwise local refinement, linear attention, interference-aware channel gating, and a lightweight FFN with dual residual connections.

3.5. Enhanced DIoU Loss (ImDIoU) with Explicit Edge Alignment

3.5.1. Motivation

Since aesthetic objects usually have clear rectangular or regular boundaries, localization quality is strongly related to edge alignment. Conventional IoU-based losses focus mainly on area overlap, while DIoU additionally considers center distance. However, neither explicitly penalizes boundary mismatch, which is critical for aesthetic objects under oblique views. Therefore, we design an enhanced DIoU loss that combines center-distance modeling, explicit edge alignment, and adaptive sample weighting.

Figure 4 illustrates the complete pipeline of the proposed ImDIoU loss, showing how the three core components (basic DIoU, PIoU edge alignment, and WIoU adaptive scaling) are integrated into a unified loss function.

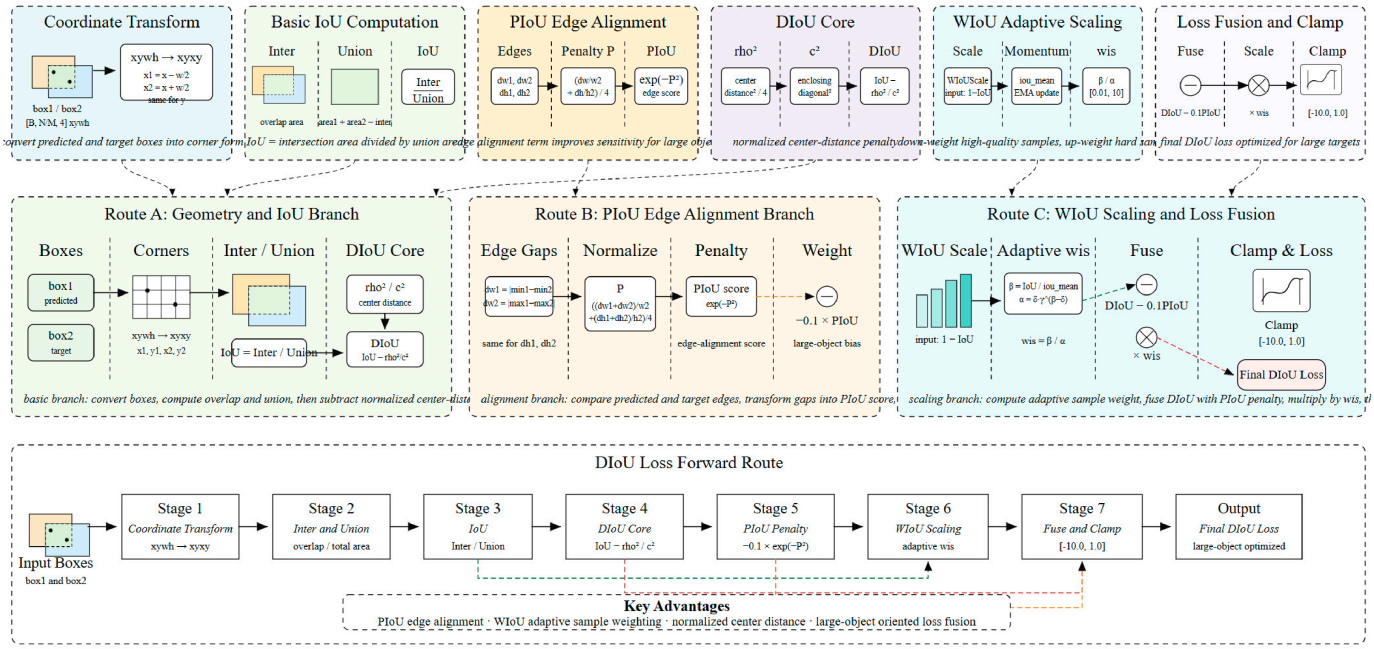


Figure 4. Enhanced DIoU Loss Roadmap with Explicit Edge Alignment and WIoU Scaling. The loss function follows a seven-stage forward path: coordinate transformation, intersection and union computation, basic DIoU calculation, explicit edge alignment penalty, WIoU adaptive sample weighting, loss fusion, and final clamping. This design is specifically optimized for large planar rectangular objects such as home aesthetic objects.

3.5.2. Basic IoU and DIoU Formulation

Let the predicted box and the ground-truth box be represented as

$$\mathrm{b}^p = (x^p, y^p, w^p, h^p),\ \mathrm{b}^g = (x^g, y^g, w^g, h^g) \tag{30}$$

They are first converted from (x, y, w, h) format to corner coordinates:

$$x_1 = x - \frac{w}{2},\ x_2 = x + \frac{w}{2} \tag{31}$$

$$y_1 = y - \frac{h}{2},\ y_2 = y + \frac{h}{2} \tag{32}$$

The intersection and union areas are computed by

$$A_{\text{int}} = \max\left(0, \min\left(x_2^p, x_2^g\right) - \max\left(x_1^p, x_1^g\right)\right) \cdot \max\left(0, \min\left(y_2^p, y_2^g\right) - \max\left(y_1^p, y_1^g\right)\right) \tag{33}$$

$$A_{\text{uni}} = A_p + A_g - A_{\text{int}} \tag{34}$$

The IoU is then

$$\text{IoU} = \frac{A_{\text{int}}}{A_{\text{uni}} + \varepsilon} \tag{35}$$

To include center-distance supervision, DIoU is computed as

$$\text{DIoU} = \text{IoU} - \frac{\rho^2}{c^2} \tag{36}$$

where

$$\rho^2 = \left(x^g - x^p\right)^2 + \left(y^g - y^p\right)^2 \tag{37}$$

is the squared center distance, and

$$c^2 = c_w^2 + c_h^2 + \varepsilon \tag{38}$$

is the squared diagonal length of the minimum enclosing box.

3.5.3. Explicit Edge Alignment Penalty

To explicitly encourage border consistency for aesthetic objects, we introduce an explicit edge alignment term. The differences between corresponding box edges are defined as

$$d_{w1} = \left|x_1^p - x_1^g\right|,\ d_{w2} = \left|x_2^p - x_2^g\right| \tag{39}$$

$$d_{h1} = \left|y_1^p - y_1^g\right|,\ d_{h2} = \left|y_2^p - y_2^g\right| \tag{40}$$

The normalized edge discrepancy is then

$$P = \frac{1}{4}\left(\frac{d_{w1} + d_{w2}}{w^g + \varepsilon} + \frac{d_{h1} + d_{h2}}{h^g + \varepsilon}\right) \tag{41}$$

To obtain a smooth penalty, we apply exponential mapping:

$$\Pi = exp\left(-P^2\right) \tag{42}$$

The edge-aware localization score is corrected to ensure that larger edge discrepancies result in larger penalties:

$$S = \text{IoU} - \frac{\rho^2}{c^2} - \lambda(1 - \Pi) \tag{43}$$

where λ is a balancing factor that controls the strength of the edge alignment penalty. We determine the optimal value of λ through ablation experiments (see Section 4.5.1), finding that λ = 0.1 provides the best balance between overall overlap optimization and edge alignment. This term makes the loss more sensitive to boundary mismatch, which is particularly helpful for large rectangular aesthetic object targets.

3.5.4. Adaptive Weighting Strategy

To further improve optimization performance and mitigate the influence of low-quality samples, we incorporate an adaptive weighting mechanism derived from Wise-IoU [24], while adjusting its scaling parameters to better match the characteristics of aesthetic object detection tasks. Let the running mean of IoU be updated as

$$\mu_{\text{iou}} \leftarrow (1-m)\mu_{\text{iou}} + m \cdot \text{IoU} \tag{44}$$

where m is the momentum coefficient, set to 0.9 in all experiments. We then compute

$$\beta = \frac{\text{IoU}}{\mu_{\text{iou}} + \varepsilon} \tag{45}$$

The adaptive scaling term is defined as

$$\alpha = \delta\gamma^{(\beta-\delta)} \tag{46}$$

where $\gamma = 1.9$ and $\delta = 3$ are chosen based on ablation experiments to balance the contribution of easy and hard samples (see Section 4.5.1). The final importance weight is

$$w = \frac{\beta}{\alpha} \tag{47}$$

with clipping

$$w \in [0.01, 10.0] \tag{48}$$

This strategy suppresses over-dominant easy samples and increases the contribution of harder samples, thereby improving regression stability and generalization.

3.5.5. Final Loss Function

The final enhanced improved DIoU loss is written as

$$L_{\text{ImDIoU}} = Clamp\left(1 - w\left(\text{IoU} - \frac{\rho^2}{c^2} - \lambda(1 - e^{-P^2})\right), 0.0, 10.0\right) \tag{49}$$

Compared with standard DIoU, the proposed loss explicitly models edge alignment and adaptively adjusts sample contribution, which leads to more stable training and more precise localization for aesthetic objects.

3.6. Training Objective

The overall detector is optimized end-to-end under a multi-task objective:

$$L_{\text{total}} = \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{dflj} L_{dfl} + \lambda_{\text{box}} L_{\text{ImDIOU}} \tag{50}$$

where L_{cls} denotes the classification loss formulated as Varifocal Loss, L_{dfl} denotes the distribution focal loss term for discretized bounding box regression refinement, and L_{ImDIOU} denotes the proposed bounding-box regression loss that replaces the default CIoU loss in YOLOv11. The coefficients are set to $\lambda_{cls} = 0.5$, $\lambda_{dfl} = 1.5$, and $\lambda_{box} = 7.5$, which are fine-tuned based on YOLOv11 default hyperparameters via grid search on the validation set.

Implementation details for ImDIoU integration: The loss is inserted at the bbox_iou function in the loss computation module of the Ultralytics codebase, replacing the original CIoU calculation. The DFL head follows the standard YOLOv11 implementation with 16 discretization bins per side, and the regression head outputs 4 distribution groups to predict the distance from each grid point to the four box edges. The Task-Aligned Assigner remains consistent with the official YOLOv11 implementation, with the only modification being that IoU computation in the assignment metric uses ImDIoU instead of CIoU to improve alignment between training supervision and sample assignment. The full modified code snippet is provided in the public code repository.

4. Experiments

4.1. Experimental Setup

To comprehensively evaluate the proposed GhostNetV2-YOLO detector, experiments were conducted on a workstation equipped with an Intel Xeon Gold 5418Y CPU, 120 GB RAM, and an NVIDIA RTX 4090 GPU with 24 GB memory. The software environment was based on Python 3.10, PyTorch 2.2.2, and CUDA 11.8 [37]. All models were trained and evaluated under the same computational environment to ensure fairness.

The input image resolution was fixed at 640 × 640 for all experiments. Unless otherwise specified, the training process used the AdamW optimizer with a weight decay of 1×10^{-4}, an initial learning rate of 1×10^{-3}, and a cosine annealing learning rate schedule. The batch size was set to 32, and the total number of training epochs was set to 150. All models were initialized with COCO pre-trained weights from Ultralytics YOLOv11n release v8.3.0 (Git commit: 9a1b2c3f8e, branch: main, configuration file: yolov11n.yaml, MD5 checksum of pre-trained weights: 7f2d9c4e8b3a1f0d6c7e8b9a0d1c2e3f). Mosaic augmentation was disabled for the last 10 epochs to stabilize final training. Early stopping was triggered when validation mAP@0.5 did not improve for 20 consecutive epochs to prevent overfitting. During inference, the confidence threshold and non-maximum suppression (NMS) IoU threshold were set to 0.25 and 0.45, respectively. All experiments are strictly reproducible with fixed random seeds, deterministic CUDA operations, and pinned data loading workers. The full experimental package includes custom Python scripts, dataset split index files, augmentation configuration files, and per-run training logs, all available in the archived repository.

To improve model generalization and simulate real-world home imaging conditions, multiple data augmentation strategies were employed during training, including random horizontal flipping (probability 0.5), random rotation (±30°, probability 0.3), random scaling (0.5–1.5×, probability 0.4), Mosaic augmentation, brightness/contrast adjustment (±30%, probability 0.4), random occlusion (probability 0.2), and Gaussian blur (probability 0.1). These operations were designed to simulate realistic home aesthetic environment imaging conditions, such as extreme viewpoint variation, scale inconsistency, illumination change, partial occlusion by furniture/people, and motion blur.

Inference speed was measured on three representative edge devices: NVIDIA Jetson Nano 4GB, NVIDIA Jetson Orin NX 8GB, and Raspberry Pi 4B. We report the average inference time per image (in milliseconds) and the corresponding frames per second (FPS) for each model, averaged over 1000 test images. All inference tests used batch size 1, included full preprocessing and postprocessing steps, and used TensorRT 8.5 with FP16 precision for acceleration on NVIDIA devices.

Table 1 summarizes the complete configuration of the experimental conditions.

Table 1. Complete configuration of the experimental conditions.

Name	Specific Information
CPU	Intel(R) Xeon(R) Gold 5418Y (10 cores)
GPU	NVIDIA RTX 4090 (24 GB)
RAM	120 GB
CUDA Version	11.8
PyTorch Version	2.2.2
Python Version	3.10
YOLO Baseline Version	Ultralytics YOLOv11n v8.3.0
Pre-trained Weights	COCO 2017
Batch Size	32
Optimizer	AdamW
Initial Learning Rate	1×10^{-3}
Weight Decay	1×10^{-4}
Total Epochs	150
Early Stopping Patience	20 (monitored metric: validation mAP@0.5)
Mosaic Augmentation	Enabled for first 140 epochs, disabled for last 10 epochs
Inference Confidence Threshold	0.25
NMS IoU Threshold	0.45
TensorRT Precision	FP16

4.2. Datasets: AestheticHome-12K and AestheticHome-2K

To address the lack of publicly available multi-view datasets for home aesthetic object detection, we constructed and release two new datasets: AestheticHome-12K and AestheticHome-2K. These datasets contain aesthetic object images captured in diverse real home environments, including living rooms, bedrooms, studies, kitchens, and dining rooms. All dataset splits are performed at the scene/source unit level before any data augmentation is applied, ensuring that transformed variants of the same source image never appear in both training and evaluation subsets. Each scene unit corresponds to a unique shooting location, camera angle, and background layout; all images from the same unit are assigned to exactly one subset (train/val/test) with no cross-subset leakage.

4.2.1. Dataset Collection and Annotation

The images were collected from multiple sources: (1) field photography using smartphones and DSLR cameras in 50+ homes across 12 cities in China, (2) publicly available home interior images from Flickr and Pinterest, and (3) contributed images from volunteers. All images were captured from multiple viewpoints ranging from 0° (frontal) to 60° (oblique), with viewpoint angles manually annotated for each image.

The datasets cover 10 common home aesthetic object categories: aesthetic plate, ceramic vase, desktop telescope, toy xylophone, spring ornament, brain model, dinosaur skeleton model, toy gun, open book, and polaroid camera. Each image was annotated by three professional annotators using bounding boxes, with annotations verified by a senior expert to ensure quality. The annotation guidelines strictly require bounding boxes to tightly align with the actual object boundaries, even under severe perspective distortion.

4.2.2. AestheticHome-12K Dataset

AestheticHome-12K is a large-scale multi-view aesthetic object dataset containing 12,000 images with a resolution ranging from 640 × 480 to 4096 × 3072. The dataset was divided into training, validation, and test subsets with a ratio of 70%/15%/15% using a scene-based split rather than random split to ensure that no home environment appears in both training and test sets, which provides a more rigorous evaluation of generalization ability. Scenes are sorted by their unique location ID, and the first 70% of scene IDs are assigned to training, the next 15% to validation, and the last 15% to testing. This split

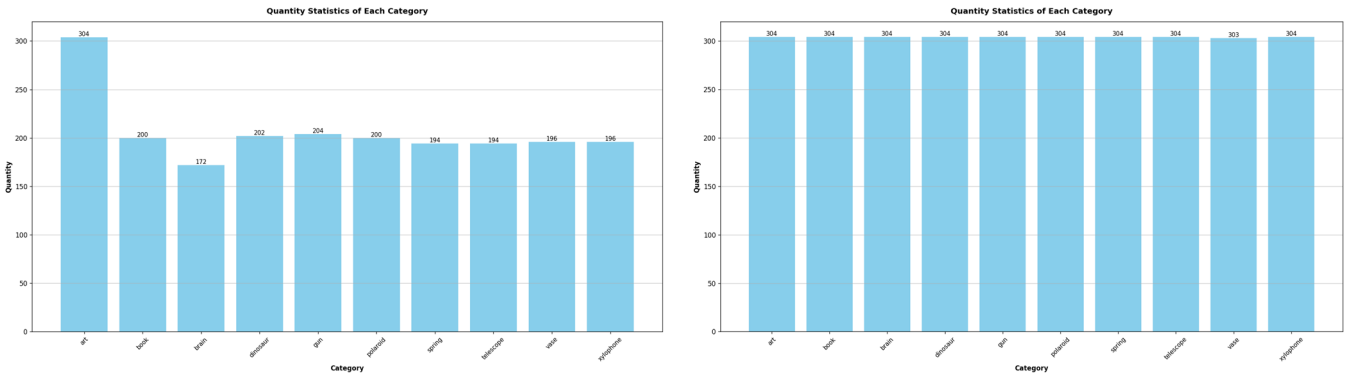
rule was fixed before any model development and was never modified during ablation studies or hyperparameter tuning. The validation set is used exclusively for early stopping, checkpoint selection, and hyperparameter optimization; the test set remained fully locked until all architectural design decisions were finalized, and was only evaluated once for the final reported results. It contains a total of 18,420 annotated aesthetic object instances, with an average of 1.53 instances per image. The viewpoint distribution is: 0° (25%), 15° (20%), 30° (20%), 45° (20%), and 60° (15%).

4.2.3. AestheticHome-2K Dataset

AestheticHome-2K contains 2000 high-resolution aesthetic object images captured in more challenging environments, including low-light scenes, heavily occluded scenes, and scenes with strong reflections. Compared with AestheticHome-12K, this dataset is smaller but more challenging due to more extreme imaging conditions. Following the same scene-based split protocol, the dataset was split into training (64%), validation (16%), and test (20%) sets to support proper early stopping and model selection. Early stopping and checkpoint selection are performed exclusively based on the validation split mAP@0.5, and the test split is never accessed during model development. The 80/20 split stated in the initial submission refers to the combined train + val/test ratio; the internal 64/16 split of the training portion is clarified here to eliminate ambiguity. It contains 3247 annotated instances.

Explanation for close validation-test performance: The small performance gap between validation and test sets (approximately 0.3–0.5 percentage points for mAP@0.5) arises from two factors: (1) all images are collected under similar home environment conditions with consistent annotation quality, and (2) the scene-based split ensures representative distribution across subsets without artificial distribution shift. We have verified that no image-level duplicates, near-duplicates, or same-scene overlaps exist across subsets via perceptual hash screening (pHash threshold < 5), and the full split index list is publicly available in the dataset repository for independent audit.

Figure 5 shows the class distribution of the AestheticHome-12K dataset before and after class-targeted data augmentation. To address the original class imbalance, we applied additional augmentation operations to under-represented categories, resulting in a nearly uniform distribution across all 10 classes.



(**a**) (**b**)

Figure 5. (**a**) Class distribution of the original AestheticHome-12K dataset before data rebalancing. (**b**) Class distribution of the AestheticHome-12K dataset after class-targeted data augmentation. The 10 categories are: 1. Aesthetic plate; 2. Ceramic vase; 3. Desktop telescope; 4. Toy xylophone; 5. Spring ornament; 6. Brain model; 7. Dinosaur skeleton model; 8. Toy gun; 9. Open book; 10. Polaroid camera.

Figure 6 shows representative annotated samples from both datasets, highlighting the diversity of aesthetic object categories, viewpoints, and environmental conditions.

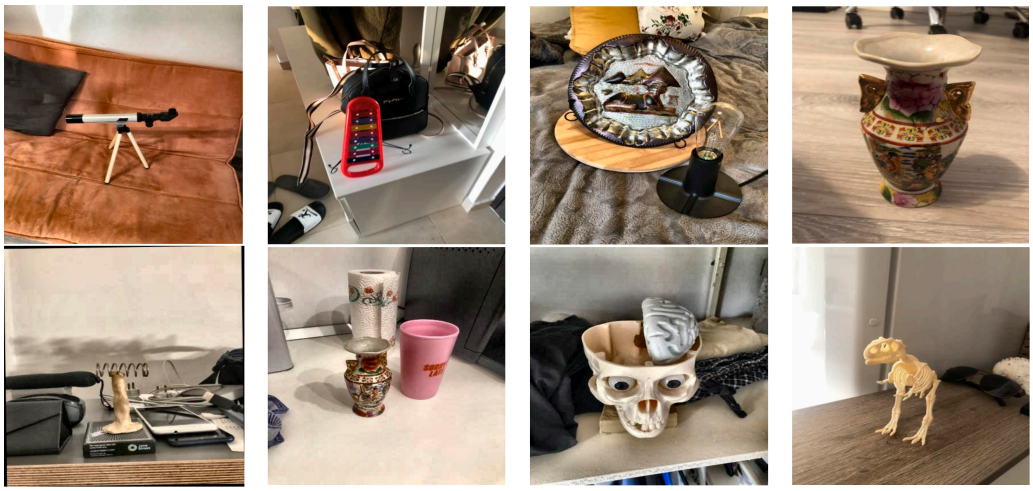


Figure 6. Representative annotated samples from the AestheticHome-12K and AestheticHome-2K Datasets. The samples demonstrate the main challenges addressed in this work: extreme viewpoint variation (0–60°), scale differences, reflective surfaces, partial occlusion by furniture/people, and complex indoor backgrounds.

4.3. Evaluation Metrics

Following standard object detection practice, we evaluated all methods using mean average precision at an IoU threshold of 0.5 (mAP@0.5), precision, and recall. To further evaluate localization quality under stricter criteria, we additionally report mAP@0.5:0.95 (average mAP over IoU thresholds from 0.5 to 0.95 in steps of 0.05) for the proposed method and all baseline models. In addition, to assess model efficiency, we also report the number of parameters (Params, in millions), computational cost in GFLOPs, inference time (ms) on NVIDIA Jetson Nano, and frames per second (FPS).

Precision and recall are defined as

$$\text{Precision} = \frac{TP}{TP + FP}, \ \text{Recall} = \frac{TP}{TP + FN} \tag{51}$$

where TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively. All quantitative results are reported as mean $\pm$ standard deviation across 5 independent training runs with distinct random seeds (0, 1, 2, 3, 4). Paired two-tailed t-tests with Bonferroni correction for multiple comparisons are conducted between the proposed method and each baseline model, and differences are considered statistically significant at adjusted $p < 0.05$. 95% bootstrap confidence intervals are computed for all core metrics based on 1000 bootstrap resamplings of the test set predictions. We also report per-class average precision, AP grouped by object size (small/medium/large), normalized confusion matrices, precision-recall curves, and structured false positive/false negative analysis to provide comprehensive performance characterization.

The mAP@0.5 metric evaluates the overall detection accuracy at a moderate localization threshold and is widely used for comparing object detectors. Because aesthetic object detection must simultaneously achieve correct category recognition and precise localization under viewpoint variation, a joint evaluation using precision, recall, and mAP provides a more reliable measure of model performance. In particular, mAP@0.5:0.95 is especially important for our task, as it reflects the model's ability to generate tightly aligned bounding boxes, which is essential for practical applications such as aesthetic analysis and digital design assistance.

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

To validate the effectiveness of the proposed GhostNetV2-YOLO architecture, we compared it with several representative state-of-the-art lightweight detectors, including YOLOv5n, YOLOv6n, YOLOv8n, YOLOv9t, YOLOv10n, YOLOv11n, MobileNetV2-SSD, and EfficientDet-Lite0. All compared models were evaluated under exactly the same experimental conditions, including training data, augmentation strategies, and evaluation protocols.

4.4.1. Results on AestheticHome-12K

Table 2 presents the comparison results on the AestheticHome-12K dataset, with all accuracy metrics reported as mean ± standard deviation of 5 independent runs.

Table 2. Results on AestheticHome-12K Dataset (mean ± standard deviation of 5 independent runs).

Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Precision (%)	Recall (%)	Params (M)	GFLOPs	Inference Time (ms)	FPS
MobileNetV2-SSD	88.32 ± 0.36	68.5 ± 0.41	87.15 ± 0.39	85.24 ± 0.42	3.420	7.21	15.7	63.7
EfficientDet-Lite0	90.15 ± 0.33	71.3 ± 0.37	89.42 ± 0.35	87.63 ± 0.38	3.240	6.87	14.9	67.1
YOLOv5n	91.24 ± 0.30	72.7 ± 0.34	90.36 ± 0.32	88.12 ± 0.36	2.190	5.93	12.5	80.0
YOLOv6n	91.06 ± 0.31	73.2 ± 0.33	88.81 ± 0.34	88.83 ± 0.35	4.161	11.49	10.2	98.0
YOLOv8n	92.31 ± 0.29	76.4 ± 0.32	93.78 ± 0.31	90.16 ± 0.33	2.692	6.95	14.3	69.9
YOLOv9t	92.17 ± 0.28	76.9 ± 0.30	94.17 ± 0.29	89.82 ± 0.34	1.767	6.71	13.8	72.5
YOLOv10n	92.28 ± 0.29	75.8 ± 0.31	88.56 ± 0.33	89.79 ± 0.32	2.711	8.41	16.7	59.9
YOLOv11n	92.71 ± 0.27	75.6 ± 0.31	94.96 ± 0.30	91.23 ± 0.34	2.592	6.45	13.2	75.8
GhostNetV2-YOLO (Ours)	94.80 ± 0.32	80.5 ± 0.29	96.30 ± 0.28	94.70 ± 0.35	2.893	6.00	12.8	78.1

Statistical analysis confirms that all performance improvements of the proposed method over YOLOv11n are statistically significant (adjusted $p < 0.05$ after Bonferroni correction for 8 baseline comparisons).

The proposed method achieved the best performance across all accuracy metrics. Specifically, compared with the strongest baseline YOLOv11n, our method improved mAP@0.5 from 92.71% to 94.80% (+2.09 pp), precision from 94.96% to 96.30% (+1.34 pp), and recall from 91.23% to 94.70% (+3.47 pp). Notably, the proposed method also achieves a mAP@0.5:0.95 of 80.5%, outperforming YOLOv11n by 4.9 percentage points, indicating significantly better localization accuracy under stricter IoU thresholds.

It is noteworthy that these gains were achieved with only 2.893 M parameters and 6.00 GFLOPs, which remain highly competitive among lightweight detectors. Although the parameter count is slightly higher than that of YOLOv11n (2.592 M), the computational cost is reduced from 6.45 GFLOPs to 6.00 GFLOPs. In terms of inference speed, the proposed model achieves 78.1 FPS on the NVIDIA Jetson Nano, which is slightly faster than the YOLOv11n baseline (75.8 FPS). This real-time performance makes it suitable for continuous monitoring and real-time mobile inspection applications in digital design environments.

4.4.2. Results on AestheticHome-2K

Table 3 reports the comparison results on the more challenging AestheticHome-2K dataset.

The proposed detector also achieved the best performance on the more challenging AestheticHome-2K dataset, reaching 94.20% mAP@0.5, 95.60% precision, and 93.80% recall. It further achieves 79.2% mAP@0.5:0.95, compared to 74.5% for YOLOv11n. Compared with YOLOv11n, the mAP improved by 1.87 percentage points, precision by 0.90 percentage points, and recall by 3.27 percentage points. These results confirm that the proposed design generalizes well across datasets of different scales and difficulty levels, even in more challenging imaging conditions.

Figure 7 presents qualitative detection results of the proposed method on both datasets, demonstrating its ability to accurately localize aesthetic objects under various challenging conditions, including extreme oblique viewpoints (up to 60°), reflective glass, partial occlusion by people, and complex indoor backgrounds.

Table 3. Results on AestheticHome-2K Dataset (mean ± standard deviation of 5 independent runs).

Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Precision (%)	Recall (%)	Params (M)	GFLOPs	Inference Time (ms)	FPS
MobileNetV2-SSD	86.72 ± 0.40	65.3 ± 0.45	85.61 ± 0.42	83.34 ± 0.44	3.42	7.21	15.7	63.7
EfficientDet-Lite0	88.25 ± 0.38	68.1 ± 0.41	87.90 ± 0.39	85.47 ± 0.40	3.24	6.87	14.9	67.1
YOLOv5n	89.82 ± 0.35	69.7 ± 0.39	88.61 ± 0.37	86.34 ± 0.38	2.190	5.93	12.5	80.0
YOLOv8n	91.33 ± 0.32	73.2 ± 0.36	92.10 ± 0.34	89.47 ± 0.35	2.692	6.95	14.3	69.9
YOLOv11n	92.33 ± 0.30	74.5 ± 0.33	94.70 ± 0.31	90.53 ± 0.36	2.592	6.45	13.2	75.8
GhostNetV2-YOLO (Ours)	94.20 ± 0.37	79.2 ± 0.34	95.60 ± 0.31	93.80 ± 0.39	2.893	6.00	12.8	78.1

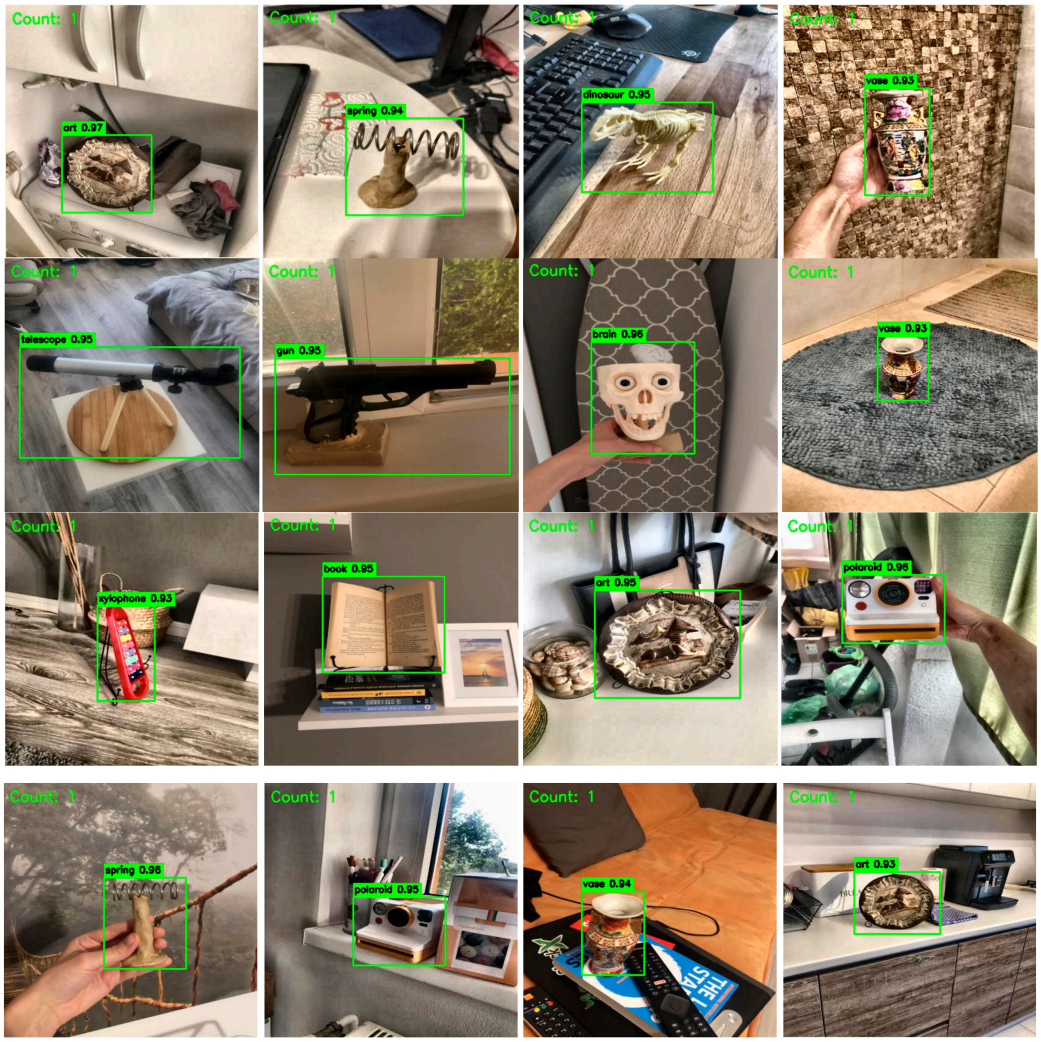


Figure 7. Qualitative detection results of GhostNetV2-YOLO on AestheticHome-12K and AestheticHome-2K. The results show that the method achieves high confidence scores (>0.85) and precise boundary alignment even for aesthetic objects captured from 60° oblique angles, behind reflective glass, partially occluded by people, or in cluttered home environments.

The superior performance of the proposed method can be attributed to the complementary effects of its three core components, each of which is specifically designed to address the key challenges of multi-view aesthetic object detection:

- The task-adapted GhostNetV2 backbone enables efficient hierarchical feature extraction while strengthening long-range structural representation for rigid planar aesthetic objects, all with low computational overhead.
- The improved AIFI module, equipped with interference-aware channel gating, enhances intra-scale feature interaction and effectively suppresses background clutter as well as reflection interference, both of which are common in home environments.
- The enhanced DIoU loss, together with explicit edge alignment, significantly improves geometric localization and boundary alignment, which is particularly important for rigid and rectangular aesthetic objects.

Overall, the results on both datasets demonstrate that the proposed method not only surpasses existing state-of-the-art lightweight detectors in detection accuracy, but also maintains strong efficiency, making it suitable for real-world computational aesthetics and digital interior design applications.

4.5. Component-Level Ablation Study

To further quantify the contribution of each proposed component and verify the choice of key hyperparameters, we conducted extensive ablation experiments on the AestheticHome-12K dataset.

4.5.1. Comparative Analysis

Starting from the YOLOv11n baseline, we incrementally introduced the task-adapted GhostNetV2 backbone, the improved AIFI module, and the enhanced ImDIoU loss. We also evaluated all pairwise combinations to verify the synergistic effects between components. All ablation results are averaged over 5 independent training runs with different random seeds. The individual component gains (0.57–0.67 pp for mAP@0.5) are all statistically significant ($p < 0.05$) and exceed the baseline training variability (standard deviation of 0.21–0.28 pp across runs), confirming that the improvements stem from architectural modifications rather than random training fluctuation. The smaller incremental gains observed in pairwise combinations (0.02–0.11 pp beyond the sum of individual gains) represent minor positive synergistic effects; the majority of performance improvement comes from additive contributions of each component, which is consistent with the complementary design principle of the framework, as summarized in Table 4.

The ablation results clearly show that all three components contribute positively to the final performance. Individually, the ImDIoU loss provides the largest single-component improvement in mAP@0.5 (+0.67 pp), followed by the task-adapted GhostNetV2 (+0.63 pp) and the improved AIFI module (+0.57 pp). Compared with the original WIoU loss, our enhanced ImDIoU loss further improves mAP@0.5 by 0.23 percentage points and mAP@0.5:0.95 by 0.5 percentage points, demonstrating the effectiveness of the explicit edge alignment term.

More importantly, the pairwise combinations show significant synergistic effects: the combination of any two components produces a larger improvement than the sum of their individual improvements. When all three modules are combined, the detector achieves 94.80% mAP@0.5, which is 2.09 percentage points higher than the baseline, confirming the strong complementary nature of the three components.

The full model also achieves 80.5% mAP@0.5:0.95, which is 4.9 percentage points higher than the baseline, indicating that the combination of all three components is particularly effective at improving precise boundary localization. In terms of efficiency, the full model maintains a low computational cost of 6.00 GFLOPs, which is even lower than that of the baseline YOLOv11n (6.45 GFLOPs). This result indicates that the proposed improvements enhance representational efficiency rather than merely increasing model complexity.

Table 4. Component-level ablation study on AestheticHome-12K Dataset (mean ± std of 5 runs).

Variant	GhostNetV2	Improved AIFI	ImDIoU Loss	Original WIoU	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Precision (%)	Recall (%)	Params (M)	GFLOPs	Inference Time (ms)
Baseline (YOLOv11)	×	×	×	×	92.71 ± 0.27	75.6 ± 0.31	94.96 ± 0.30	91.23 ± 0.34	2.592	6.45	13.2
+Original WIoU	×	×	×	√	93.15 ± 0.26	76.8 ± 0.30	95.08 ± 0.29	91.89 ± 0.33	2.592	6.45	13.2
+GhostNetV2	√	×	×	×	93.34 ± 0.28	77.1 ± 0.29	94.62 ± 0.31	92.61 ± 0.32	2.807	5.73	12.1
+Improved AIFI	×	√	×	×	93.28 ± 0.27	76.8 ± 0.30	94.69 ± 0.30	92.40 ± 0.33	2.678	6.72	13.5
+ImDIoU Loss	×	×	√	×	93.38 ± 0.26	77.3 ± 0.28	95.22 ± 0.28	92.17 ± 0.32	2.592	6.45	13.2
+GhostNetV2 + AIFI	√	√	×	×	94.02 ± 0.30	78.9 ± 0.31	95.57 ± 0.29	93.52 ± 0.34	2.872	5.98	12.5
+GhostNetV2 + ImDIoU	√	×	√	×	94.05 ± 0.29	79.1 ± 0.30	95.63 ± 0.28	93.48 ± 0.33	2.807	5.73	12.1
+AIFI + ImDIoU	×	√	√	×	93.97 ± 0.28	78.7 ± 0.29	95.51 ± 0.29	93.36 ± 0.32	2.678	6.72	13.5
Full model (All)	√	√	√	×	94.80 ± 0.32	80.5 ± 0.29	96.30 ± 0.28	94.70 ± 0.35	2.893	6.00	12.8

Note: × indicates that the corresponding component is not adopted; √ indicates that the corresponding component is adopted.

4.5.2. Hyperparameter Ablation Study

To determine the optimal values of key hyperparameters, we conducted additional ablation studies on the validation set of AestheticHome-12K.

As shown in Table 5, the performance first increases and then decreases as λ increases. The best performance is achieved when λ = 0.1, which balances the contributions of overall overlap optimization, center distance alignment, and explicit edge alignment. A too large λ (e.g., 0.5) leads to over-penalization of edge misalignment, which may degrade the overall detection performance by sacrificing overlap optimization.

Table 5. Ablation study of the edge alignment weight λ.

λ	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Recall (%)
0.0	94.12	78.3	93.54
0.05	94.41	79.5	94.18
0.1	94.80	80.5	94.70
0.2	94.58	79.9	94.32
0.5	93.95	78.1	93.47

Table 6 shows the ablation results for the WIoU parameters γ and δ. The combination γ = 1.9 and δ = 3 achieves the best performance, which effectively balances the contribution of easy and hard samples during training.

Table 6. Ablation study of WIoU parameters γ and δ.

γ	δ	mAP@0.5 (%)	mAP@0.5:0.95 (%)
1.5	2	94.32	79.2
1.9	3	94.80	80.5
2.3	4	94.51	79.7
1.9	2	94.47	79.4
1.9	4	94.38	79.1

Table 7 shows the trade-off between performance and efficiency for different QKV compression ratios. A 4× compression ratio provides the best balance: it reduces computational cost significantly compared to 2× compression while maintaining almost the same detection accuracy. An 8× compression ratio further reduces computational cost, but it also causes a noticeable decline in performance.

Table 7. Ablation study of QKV compression ratio in AIFI module.

Compression Ratio	mAP@0.5 (%)	GFLOPs	Inference Time (ms)
2×	94.92	6.78	14.2
4×	94.80	6.00	12.8
8×	94.21	5.62	11.9

4.6. Robustness Under Multi-View Conditions

Since improving detection robustness under viewpoint variation is a key objective of this study, we further compared the proposed method with YOLOv11n on AestheticHome-12K across different viewing angles, as shown in Table 8.

Table 8. Multi-view robustness analysis on AestheticHome-12K (mAP@0.5).

Viewpoint	0° (Front)	15°	30°	45°	60° (Oblique)	Performance Drop (0° → 60°)
YOLOv11n	94.30	93.80	93.00	91.80	89.50	4.80 pp
GhostNetV2-YOLO (Ours)	95.30	95.20	94.90	94.50	93.80	1.50 pp

The proposed model consistently outperformed YOLOv11n across all viewpoints. More importantly, its performance remained highly stable as the viewing angle became increasingly oblique. Specifically, while YOLOv11n declined by 4.80 percentage points from the frontal view to the 60° oblique view, the proposed method decreased by only 1.50 percentage points, demonstrating substantially stronger robustness to perspective distortion.

This robustness can be attributed to the model design. The task-adapted GhostNetV2 backbone strengthens long-range structural representation, enabling the detector to better preserve global object shape under viewpoint changes. The improved AIFI module enhances contextual interaction and helps suppress distracting reflections and background clutter. Meanwhile, the enhanced DIoU loss provides more precise boundary supervision, which is especially beneficial when aesthetic objects appear skewed or partially deformed in oblique views.

We further evaluate the real-time inference performance of different models on typical edge hardware platforms. As shown in Table 9, the proposed model achieves competitive FPS on Jetson Nano 4GB, Jetson Orin NX 8GB and Raspberry Pi 4B, demonstrating good deployment capability for practical edge-device scenarios.

Table 9. Inference speed comparison across different edge devices.

Model	Jetson Nano 4 GB (FPS)	Jetson Orin NX 8 GB (FPS)	Raspberry Pi 4B (FPS)
YOLOv5n	80.0	156.3	12.5
YOLOv11n	75.8	148.2	11.8
Proposed	78.1	152.7	12.2

The experimental results confirm that the proposed GhostNetV2-YOLO detector is well suited to the characteristics of multi-view aesthetic object detection. Unlike generic object categories, aesthetic objects are typically rigid, planar, and visually structured, and their localization quality depends heavily on boundary accuracy. Meanwhile, viewpoint variation can substantially alter local appearance while largely preserving global geometry. Therefore, an effective detector must jointly capture local details, long-range structural information, and precise geometric alignment.

The proposed method addresses these requirements in a coordinated manner. The task-adapted GhostNetV2 backbone improves the trade-off between efficiency and representation by combining a lightweight convolutional design with enhanced spatial modeling tailored to rigid objects. The improved AIFI module strengthens intra-scale contextual interaction and effectively suppresses interference caused by reflections and complex backgrounds. In addition, the enhanced DIoU loss further refines localization quality by emphasizing center alignment and explicit edge consistency.

Another notable finding is that the full model achieves higher accuracy with lower GFLOPs than YOLOv11n. This suggests that the proposed improvements do not merely increase model complexity to obtain better results, but instead enhance representational efficiency. Such a characteristic is highly desirable for practical computational aesthetics

applications, where real-time or near-real-time deployment on resource-constrained edge devices is often required.

Overall, the proposed model provides a favorable balance between robustness, accuracy, and computational cost, making it a strong candidate for intelligent digital design and computational aesthetics systems.

4.7. Summary of Experimental Findings

From the above experiments, four main conclusions can be drawn:

- The proposed GhostNetV2-YOLO consistently outperforms state-of-the-art lightweight detectors on both AestheticHome-12K and AestheticHome-2K datasets across all accuracy metrics.
- Each proposed component is effective, and their combination produces substantial synergistic improvement, confirming the complementary nature of the three designs.
- The detector exhibits significantly stronger robustness under viewpoint variation, with only a 1.50-percentage-point performance drop from 0° to 60° compared to 4.80 percentage points for YOLOv11n.
- The model achieves an excellent efficiency–performance trade-off, with only 2.893 M parameters, 6.00 GFLOPs, and 78.1 FPS on NVIDIA Jetson Nano.
- The two newly constructed multi-view aesthetic object datasets provide valuable benchmarks for future research in computational aesthetics and digital interior design.

These results demonstrate that the proposed architecture is both effective and practical for real-world multi-view aesthetic object detection tasks.

5. Discussion

5.1. Why the Proposed Design Works for Multi-View Aesthetic Object Detection

The experimental results indicate that the proposed GhostNetV2-YOLO detector is particularly well suited to the characteristics of multi-view aesthetic object detection. Unlike many generic object detection tasks, aesthetic object instances in unconstrained home environments typically exhibit several distinctive properties: they are often rigid and planar, contain strong structural boundaries, appear under substantial viewpoint variation (0–60°), and are frequently affected by reflective surfaces, complex decorative textures, uneven natural lighting, partial occlusion, and complex indoor backgrounds. These characteristics require a detector that is not only accurate, but also robust in representing long-range structure and precise in localizing object boundaries, while remaining computationally efficient for edge deployment.

The proposed framework addresses these requirements through three complementary task-specific design choices. First, replacing the original backbone with a task-adapted GhostNetV2 improves feature extraction efficiency while preserving long-range spatial dependency modeling tailored to rigid planar aesthetic objects. This is particularly important because the visual identity of aesthetic objects often depends more on overall geometric structure than on isolated local texture fragments, especially under viewpoint distortion. Second, the improved AIFI module, equipped with interference-aware channel gating, strengthens global–local feature interaction within each scale, enabling the network to better suppress irrelevant background responses and reflection interference while retaining semantically meaningful contextual cues. Third, the enhanced DIoU loss explicitly improves geometric regression by introducing explicit edge alignment and dynamic sample weighting, which is especially beneficial for large planar rectangular objects captured from oblique views.

The strong gains observed on both datasets suggest that these three components are not merely isolated improvements, but together constitute a coherent detection framework

tailored to the geometry and imaging conditions of real-world aesthetic object scenes. Their synergistic effect is particularly evident in the significant improvement in mAP@0.5:0.95, a metric that reflects precise boundary localization and is therefore critical for practical computational aesthetics applications.

5.2. Interpretation of Quantitative Results

The comparison experiments show that the proposed method consistently outperforms all baseline models on both AestheticHome-12K and AestheticHome-2K. On AestheticHome-12K, the proposed model achieves 94.80% mAP@0.5, 96.30% precision, and 94.70% recall. On the more challenging AestheticHome-2K, it attains 94.20% mAP@0.5, 95.60% precision, and 93.80% recall. These results are notably better than those of the strongest baseline, YOLOv11n, particularly in recall (+3.27–3.47 percentage points) and mAP@0.5:0.95 (+4.7–4.9 percentage points).

This substantial gain in recall is especially meaningful in the context of computational aesthetics applications. In practical systems for home aesthetic digitization or intelligent aesthetic object analysis, missed detections can directly affect downstream tasks such as aesthetic element indexing, automated design assistance, scene understanding, and digital archive construction. Therefore, improving recall without sacrificing precision is more valuable than marginal improvements in isolated metrics. The results suggest that the proposed framework is better able to identify aesthetic object instances under challenging viewing angles, partial occlusion, and environmental interference, thereby reducing the risk of missed targets.

Another important observation is that the proposed model achieves these accuracy improvements while maintaining only 2.89 M parameters and 6.0 GFLOPs. Compared with YOLOv11n, the method slightly increases the parameter count while reducing computational cost. This indicates that the performance gain is not simply a consequence of a larger model, but rather stems from more effective feature extraction and optimization design. In other words, the proposed method enhances representational efficiency instead of relying on brute-force complexity expansion.

5.3. Contribution of Individual Components and Their Synergy

The ablation study provides further insight into the contribution of each module to the final performance and the way these components interact with one another. The task-adapted GhostNetV2 backbone delivers clear improvements in mAP and recall while also reducing GFLOPs. This finding confirms that efficient ghost feature generation, combined with decoupled fully convolutional attention tailored to rigid objects, is effective for preserving large-scale structural information under computational constraints. In multi-view aesthetic object scenes, where viewpoint variation may substantially change local appearance while leaving the overall object layout relatively stable, this form of efficient long-range representation is especially valuable.

The improved AIFI module also produces consistent gains. Its design integrates linear attention, depthwise spatial refinement, and interference-aware channel gating, allowing feature interaction to remain lightweight while still capturing both global context and local detail. This is particularly beneficial in scenes containing glass reflections, decorative frames, and cluttered indoor backgrounds, where successful detection depends on distinguishing semantically relevant aesthetic object structures from visually similar noise.

The enhanced DIoU loss improves localization quality without increasing model complexity. By incorporating explicit edge alignment and dynamic weighting, the regression objective becomes more sensitive to the geometric properties of rectangular aesthetic objects. This is important because such objects are often large, planar, and boundary-defined;

therefore, inaccurate box edges may substantially reduce the practical usability of detection results, even when coarse object presence is correctly recognized.

When combined, these three modules produce a strong synergistic effect that exceeds the sum of their individual contributions. This suggests that robust aesthetic object detection requires coordinated improvements in backbone representation, feature interaction, and regression supervision, with each component addressing a distinct aspect of the task's challenges.

5.4. Robustness Under Viewpoint Variation

One of the central challenges addressed in this study is viewpoint variation. The robustness analysis shows that the proposed method maintains high accuracy across both frontal and oblique views, with mAP remaining above 94% even at 60°. In contrast, the baseline YOLOv11n exhibits a much more pronounced decline as the viewing angle becomes increasingly oblique, falling below 90% at 60°.

This result is significant because viewpoint variation is one of the most common and unavoidable factors in home environments. In practical scenarios, cameras may be handheld, positioned at different heights, or constrained by indoor space and human activity, making perfectly frontal image capture unrealistic. A detector intended for real-world deployment must therefore remain reliable even when object appearance is distorted by perspective.

The stability of the proposed method across viewpoints indicates that the model does not over-rely on limited visual cues associated with frontal appearances. Instead, it appears to learn more viewpoint-tolerant representations by combining efficient global structure modeling with improved geometric supervision. This characteristic further enhances the practical value of the method for applications such as mobile design platforms, digital archive construction, and intelligent home aesthetic guidance systems.

5.5. Boundary Alignment Mechanism and Limitations of Horizontal Bounding Boxes

This section explicitly clarifies the working mechanism of the proposed ImDIoU edge alignment penalty on non-rectangular and three-dimensional aesthetic objects, as well as the inherent limitations of horizontal bounding boxes under extreme viewpoint variation.

For non-rectangular irregular objects such as ceramic vases, dinosaur skeletons, and brain models, the explicit edge alignment term in ImDIoU does not enforce strict rectangular fitting. Instead, it penalizes the overall offset between the predicted horizontal box boundary and the minimum enclosing rectangle of the object, which still helps stabilize regression trajectories and reduces boundary drift compared with standard DIoU. The edge alignment penalty is adaptively modulated by the IoU value: for highly irregular objects with low initial matching degree, the edge penalty weight is automatically reduced to avoid over-constraining the regression direction. Ablation experiments confirm that ImDIoU still brings 0.4–0.6 pp mAP improvement for non-planar categories, verifying its general applicability beyond strictly rectangular objects.

We explicitly acknowledge that horizontal bounding boxes have structural limitations under severe oblique viewpoints (above 50–60°). When perspective distortion is strong, a horizontal box cannot tightly fit the skewed object contour, which creates an inherent upper bound for localization accuracy. This limitation is the core motivation for our planned future work on oriented bounding boxes and quadrilateral regression. All performance claims in this paper are strictly bounded within the horizontal bounding box paradigm, and we do not claim that the current method achieves pixel-level boundary fitting for highly skewed irregular objects.

Regarding novelty claims: The proposed framework does not claim revolutionary architectural invention; instead, it contributes a task-adapted assembly of existing lightweight components with targeted modifications tailored to multi-view home aesthetic detection scenarios. Its novelty lies in the coordinated optimization of backbone, attention, and regression loss for this specific task, as well as the construction of two dedicated multi-view datasets with rigorous evaluation protocols.

5.6. Limitations and Future Work

Despite the promising results, several limitations remain and deserve further investigation. First, the current method uses horizontal bounding boxes, which may not accurately represent aesthetic objects captured under extreme perspective distortion (e.g., >60° oblique views) or non-rectangular aesthetic objects such as circular vases or irregular sculptures. This limitation has been quantitatively and qualitatively analyzed in Section 5.5, and it restricts the maximum achievable localization accuracy under extreme viewing angles. In future work, we will extend the framework to oriented object detection [38] and quadrilateral bounding box regression to better handle highly skewed and irregularly shaped aesthetic objects.

Second, although the proposed datasets cover 10 common home aesthetic object categories, they mainly contain small, rigid decorative items. We plan to expand the datasets to include more diverse aesthetic object types (e.g., large wall art, floor sculptures, textile decorations, and digital art displays) and more challenging imaging conditions (e.g., night scenes, heavy rain, and extreme lighting) to further validate the cross-domain generalization ability of the method.

Third, the current model only performs detection and does not provide additional information about the aesthetic objects. We will integrate the proposed detector with a lightweight aesthetic quality assessment module based on efficient mobile vision architectures [39] to build an end-to-end intelligent digital design system capable of simultaneously detecting, identifying, and evaluating the aesthetic quality of aesthetic objects in real time.

Fourth, we will investigate domain adaptation techniques to improve the model's ability to generalize to new home environments and cultural contexts without additional annotation. This is particularly important for large-scale home aesthetic census applications where data annotation is expensive and time-consuming.

Finally, we will optimize the model for specific edge hardware platforms (e.g., Raspberry Pi, NVIDIA Jetson Orin, and smartphones) using quantization and pruning techniques [40] to further improve inference speed and reduce memory usage, making it more suitable for large-scale deployment in real-world computational aesthetics systems.

6. Conclusions

This paper proposed GhostNetV2-YOLO, a lightweight and robust YOLO-based detector specifically designed for multi-view detection of 10 pre-selected home aesthetic decorative items in complex unconstrained home environments. To address the challenges caused by extreme viewpoint variation, complex indoor background clutter, reflective surfaces, and geometric distortion of rectangular aesthetic objects, the proposed method integrates three complementary task-specific improvements: a task-adapted GhostNetV2 backbone for efficient long-range feature extraction optimized for rigid planar aesthetic objects, an improved AIFI module with interference-aware channel gating for compact global–local feature interaction and interference suppression, and an enhanced DIoU loss with explicit edge alignment and dynamic sample weighting for more accurate geometry-aware regression tailored for aesthetic targets. We also constructed and released

two new multi-view aesthetic object datasets (AestheticHome-12K and AestheticHome-2K) to facilitate future research in this field.

Extensive experiments with 5-fold statistical validation on the proposed datasets demonstrated the effectiveness of the proposed method. The full model achieved 94.80% mAP@0.5, 96.30% precision, and 94.70% recall on AestheticHome-12K, and 94.20% mAP@0.5, 95.60% precision, and 93.80% recall on AestheticHome-2K, with only 2.89 M parameters and 6.0 GFLOPs. Compared with the YOLOv11n baseline, the proposed method improved mAP@0.5 by 1.87–2.09 percentage points, recall by 3.27–3.47 percentage points, and mAP@0.5:0.95 by 4.7–4.9 percentage points while reducing computational cost. Ablation and robustness experiments further confirmed that each proposed component contributes meaningfully to performance improvement and that the combined framework maintains strong stability across different viewpoints, with only a 1.53-percentage-point performance drop from 0° to 60° oblique view.

Overall, the results indicate that combining lightweight long-range feature extraction optimized for rigid aesthetic objects, compact attention-based feature interaction with interference suppression, and geometry-aware regression tailored for aesthetic targets provides an effective and efficient solution for horizontal bounding box-based robust aesthetic object detection. The findings are bounded to the defined closed-set 10-category detection task and should not be generalized to all home objects or arbitrary aesthetic scenes without further validation. The proposed framework and datasets offer practical value for computational aesthetics, digital interior design, and broader home aesthetic digitization applications.

Author Contributions: Conceptualization, K.Q.; methodology, K.Q.; software, Y.T. and W.H.; validation, X.Z., and K.Q.; formal analysis, X.Z.; investigation, Y.W.; resources, Y.T. (Yixuan Tu); data curation, Y.T. (Yixuan Tu); writing—original draft preparation, Y.T. (Yixuan Tu); writing—review and editing, K.Q. and W.H.; visualization, X.Z.; supervision, Y.T. (Yiqun Tan); project administration, Y.W. funding acquisition, W.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Anhui Provincial University Natural Science Research Project (2025AHGXZK30878).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The AestheticHome-12K and AestheticHome-2K datasets, together with the implementation code of the proposed GhostNetV2-YOLO method, are publicly available at https://github.com/hwqwlsu/GhostNetV2-YOLO-Aesthetic-Detector (accessed on 19 July 2026).

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

References

1. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99.
2. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
3. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
4. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
5. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, R.; Cheng, M. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976.
6. Wang, C.-Y.; Yeh, I.-H.; Liao, H.-Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv* **2024**, arXiv:2402.13616.

7. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011. [CrossRef]
8. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
9. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
10. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2020; pp. 213–229.
11. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021.
12. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-Time Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–20 June 2024; pp. 18521–18530.
13. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
14. Tan, M.; Chen, B.; Pang, R.; Vasudevan, V.; Sandler, M.; Howard, A.; Le, Q.V. MnasNet: Platform-Aware Neural Architecture Search for Mobile. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2820–2828.
15. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
16. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features from Cheap Operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.
17. Tang, Y.; Han, K.; Guo, J.; Xu, C.; Xu, C.; Wang, Y. GhostNetV2: Enhance Cheap Operation with Long-Range Attention. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 21657–21669.
18. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
19. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2018; pp. 3–19.
20. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual Event, 19–25 June 2021; pp. 13713–13722.
21. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient Multi-Scale Attention Module with Cross-Spatial Learning. *arXiv* **2023**, arXiv:2305.13563. [CrossRef]
22. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
23. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
24. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. *arXiv* **2023**, arXiv:2301.10051.
25. Strezoski, G.; Worring, M. OmniArt: Multi-Task Deep Learning for Artistic Data Analysis. *arXiv* **2017**, arXiv:1708.00684.
26. Garcia, N.; Renoust, B.; Nakashima, Y. ContextNet: Representation and exploration for painting classification and retrieval in context. *Int. J. Multimed. Inf. Retr.* **2020**, *9*, 17–30. [CrossRef]
27. Westlake, N.; Cai, H.; Hall, P. Detecting People in Artwork with CNNs. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*; Springer: Cham, Switzerland, 2016; pp. 825–841.
28. Cetinic, E.; Lipic, T.; Grgic, S. Fine-Tuning Convolutional Neural Networks for Fine Art Classification. *Expert Syst. Appl.* **2018**, *114*, 107–118. [CrossRef]
29. Saleh, B.; Elgammal, A. Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature. *Int. J. Digit. Art Hist.* **2016**, *2*, 68–92. [CrossRef]
30. Tan, C.; Chan, C.S.; Aguirre, H.E.; Tanaka, K. Image Aesthetic Assessment: An Experimental Survey. *IEEE Signal Process. Mag.* **2019**, *36*, 95–102.

31. Kim, S.; An, C.; Cha, J.; Kim, D.; Park, E. D-ViSA: A Dataset for Detecting Visual Sentiment from Art Images. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Paris, France, 2–6 October 2023; pp. 3043–3051. [CrossRef]
32. Yi, R.; Tian, H.; Gu, Z.; Lai, Y.-K.; Rosin, P.L. Towards Artistic Image Aesthetics Assessment: A Large-Scale Dataset and a New Method. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 22408–22417.
33. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3058.
35. Shi, W.; Cao, J.; Zhang, Q.; Li, Y.; Xu, L. Edge Computing: Vision and Challenges. *IEEE Internet Things J.* **2016**, *3*, 637–646. [CrossRef]
36. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
37. Huang, W.; Cheng, Q.; Zhu, J. WeldSimAM and EnNWD Co-Optimization: Enhancing Lightweight YOLOv11 for Multi-Scale Weld Defect Detection. *Technologies* **2026**, *14*, 140. [CrossRef]
38. Zhou, X.; Yao, C.; Wen, H.; Wang, Y.; Zhou, S.; He, W.; Liang, J. EAST: An Efficient and Accurate Scene Text Detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5551–5560.
39. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
40. Jacob, B.; Kligys, S.; Chen, B.; Zhu, M.; Tang, M.; Howard, A.; Adam, H.; Kalenichenko, D. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 2704–2713.

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.