Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MGA-YOLO: A Surface-Guided Multi-Task Network for Industrial Defect Detection

LingXiao Zhang
Zhongyuan University of Technology
JinZhong Xu
6180@zut.edu.cn

Zhongyuan University of Technology
PeiLiang Yao
Zhongyuan University of Technology
Jing Jing Zhang
Beijing North Vehicle Group Corporation
ZhiLiang Zhang
Beijing North Vehicle Group Corporation

Research Article

Keywords: Industrial defect detection, Multi-task learning, Residual attention mechanism, Object detection, YOLOv8, Gradient balancing

Posted Date: February 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8794809/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MGA-YOLO: A Surface-Guided Multi-Task Network for Industrial Defect Detection

LingXiao Zhang[a,b], JinZhong Xu[b,*], PeiLiang Yao[b], JingJing Zhang[c], ZhiLiang Zhang[c]

[a]*School of Computer Science, Zhongyuan University of Technology, Zhengzhou, 450007, Henan Province, China*
[b]*School of Artificial Intelligence, Zhongyuan University of Technology, Zhengzhou, 450007, Henan Province, China*
[c]*Beijing North Vehicle Group Corporation, Beijing, 100072, China*

Abstract

Industrial surface defect detection is critical for manufacturing quality control. However, complex mechanical components comprising both rough casting surfaces and precision-machined surfaces pose significant challenges due to the visual similarity between casting textures and actual defects, leading to high false-alarm rates. This paper proposes MGA-YOLO, a lightweight, real-time multi-task network optimized for high-precision defect detection on machined surfaces. Built upon a YOLOv8m shared backbone to ensure computational efficiency, MGA-YOLO introduces three key innovations:(1)A Channel-Alignment Segmentation Decoder integrated with a Boundary F1 Loss to precisely delineate machined regions;(2)A Mask-Guided Residual Attention (MGRA) module that utilizes segmentation features as spatial and channel priors to dynamically enhance feature responses in regions of interest while suppressing casting background interference;(3)A Task-Specific Bias Initialization (TSBI) strategy designed to mitigate task dominance and gradient interference during the early stages of multi-task learning.

Experimental results on an industrial dataset demonstrate that MGA-YOLO achieves a 76.7% mAP@50 and 88.1% mIoU. Compared to single-task baselines, the proposed model significantly reduces false detections in non-machined areas while maintaining a real-time inference speed of 135 FPS, satisfying the stringent requirements for edge-side deployment in industrial environments.

Keywords: Industrial defect detection, Multi-task learning, Residual attention mechanism, Object detection, YOLOv8, Gradient balancing

1. Introduction

In intelligent manufacturing, automated surface defect detection is essential for ensuring product quality and fatigue life. While deep learning models like YOLO and Faster R-CNN excel in static backgrounds, they struggle with complex industrial components featuring heterogeneous surfaces. Specifically, the chaotic textures of rough casting surfaces frequently mimic defect signatures, leading to excessive false positives in standard models. We present the Mask-Guided Attention YOLO (MGA-YOLO), a framework that deeply couples semantic segmentation with object detection via spatial attention. Rather than simple task stacking, MGA-YOLO employs a dual-stream decoder sharing a YOLOv8 encoder. The segmentation branch utilizes a channel-alignment module and Boundary F1-loss to resolve machined regions. These features serve as "soft" spatial priors for the Mask-Guided Residual Attention MGAmodule in the detection branch, which dynamically suppresses non-machined area responses. Furthermore, we introduce a Bias-based Gradient Balancing strategy to stabilize multi-task convergence. MGA-YOLOachieves 76.7% mAP and 85.6% mIoU, effectively isolating casting surface interference and providing high-precision defect localization suitable for industrial environments.

2. Related Work

2.1. Deep learning–based visual inspection

Deep learning has fundamentally transformed industrial inspection, significantly enhancing adaptability to complex backgrounds and diverse defect morphologies [1]. Classic frameworks, including Faster R-

*Corresponding author: JinZhong Xu, email: 6180@zut.edu.cn

CNN [2], SSD [3], and the YOLO series [4], established the groundwork for balancing accuracy and real-time performance. Subsequent engineering optimizations [5, 6] further validated their efficacy in surface defect detection. Recently, the Transformer architecture has shifted the detection paradigm from anchor-based to query-driven [7, 8]. Representative models such as Conditional DETR V2 [9], RT-DETR [10], and Mask DINO [11] leverage global dependency modeling to improve detection robustness against environmental noise.To address the scarcity of anomalous samples in industrial scenarios, researchers have explored data-efficient strategies. For instance, Gao [12] introduced one-normal-image prompts for multi-class anomaly detection, while Gui et al. [13] utilized few-shot anomaly-driven generation. Tu et al. [14] focused on self-supervised feature adaptation for 3D data. A comprehensive survey by Liu et al. [15] systematically categorized these into reconstruction, contrastive, and generative approaches. To further reduce omission rates, two-stage frameworks [16] and advanced synthesis methods like PBAS [17] and DiffuMatting [18] have been proposed to enrich defect representation. Meanwhile, EfficientAD [19] achieved high-precision detection at millisecond-level latencies, facilitating high-speed online deployment. However, these methods often suffer from feature interference when dealing with multi-surface hybrid inputs. Their stability remains insufficient under complex lighting, and bounding-box-only outputs lack the "physical region consistency" required for localized functional areas.

2.2. Semantic and Instance Segmentation

For tasks requiring pixel-level precision, segmentation provides finer-grained defect descriptions. Mask R-CNN [20] remains a cornerstone for instance segmentation. Recent research has extended these capabilities to open-vocabulary and cross-domain scenarios: Seg4Diff [21] leverages diffusion Transformers for semantic alignment, while SCORE [22] emphasizes scene context for remote sensing. Similar trends in query structures and unified modeling are observed in video and sequence segmentation [23].Regarding foundation model adaptation, MedSeqFT [24] and SMART [25] utilize sequential fine-tuning and semantic refocusing to handle rare categories and domain gaps. In fine-grained structures, DIN [26] enhances local detail recognition through cross-image attention. The adaptation of vision foundation models like SAM and DINOv3 [27, 28, 29] has shown immense potential, while cross-modal approaches such as CGG [30] and DALNet [31] further promote weak-supervision and grounding capabilities. Furthermore, the integration of Wavelets and Transformers [32] has explored pixel-level representation in high-noise backgrounds. Despite these advances, the high computational overhead and pixel-level annotation costs of segmentation models limit their end-to-end industrial deployment, especially where real-time performance is paramount.

2.3. Research Gaps and Challenges

Despite the significant progress in detection, anomaly detection, and segmentation, a critical gap remains in handling complex industrial parts with multi-surface hybrids. First, surface interference leads to a coexistence of false positives and missed detections. Components with complex surfaces often exhibit significant visual texture variances. Existing models, typically trained under the assumption of uniform backgrounds or single materials [13, 14], frequently trigger false alarms on casting surfaces while missing defects on machined surfaces—a problem particularly acute in real-time frameworks [5]. While query-based detection offers global modeling advantages, it still requires effective spatial constraints and prior guidance to ensure physical region consistency [4, 27]. Second, there is a lack of region-level feature isolation and "logical consistency" mechanisms. Industrial inspection often demands defect detection only within specific functional zones. However, existing methods generally lack explicit modeling for different functional areas. Although segmentation models provide pixel-level region information [15, 28], their high deployment costs are prohibitive. Similarly, while anomaly detection and generative synthesis mitigate data scarcity [6, 8, 12], they do not inherently resolve the "detection-only-on-machined-surface" constraint. Foundation models and cross-modal approaches [22, 23, 24] show potential for transfer learning but face insufficient feature response and semantic mismatch in weakly textured industrial scenarios [25, 26]. To bridge these gaps, this study proposes a "Region-Segmentation-Detection" approach. By injecting the machined surface mask as a soft spatial prior into the detection branch, we achieve region-level feature suppression and attention focusing, thereby enhancing the robustness and logical consistency of defect localization under complex lighting and multi-surface hybrid conditions.

3. Methodology

3.1. Overall Architecture

To address the requirements for real-time performance and multi-task synergy in industrial scenarios,

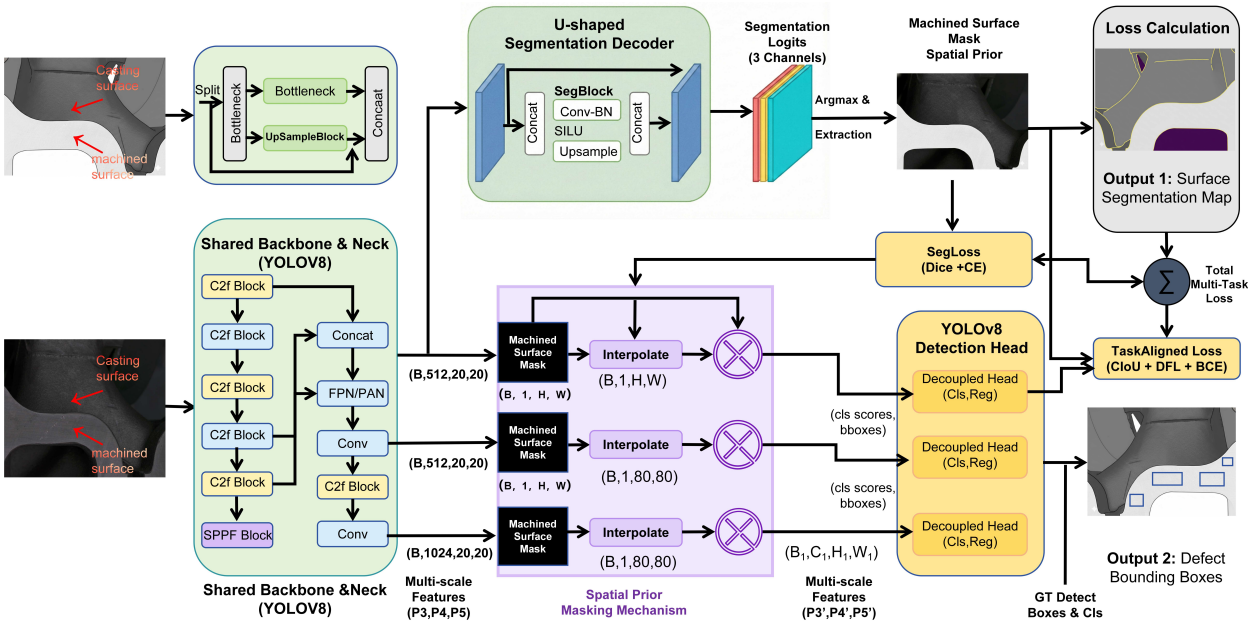


Figure 1: Overall architecture of the proposed MGA-YOLO framework.

this study proposes a multi-task joint learning network termed MGA-YOLO (Mask-Guided Attention YOLO). The network employs the feature extractor of YOLOv8 as a shared backbone. Compared to traditional architectures such as ResNet, YOLOv8 integrates C2f modules and an optimized PA-FPN (Path Aggregation Feature Pyramid Network) structure. These components provide richer gradient flows and more robust multi-scale feature representations, effectively alleviating the feature conflict issues inherent in multi-task learning. Furthermore, although the YOLO series undergoes continuous iterative updates, our empirical research on actual production line data and imaging conditions revealed that newer YOLO versions exhibit greater volatility. This instability stems from more aggressive module modifications and increasingly complex training strategies in the latest versions. Taking these factors into account, this study adopts YOLOv8 as the shared backbone and introduces minor structural modifications to the overall network. Specifically, we employ a "Shared Encoder - Decoupled Decoders"architecture, as illustrated in Figure 1.

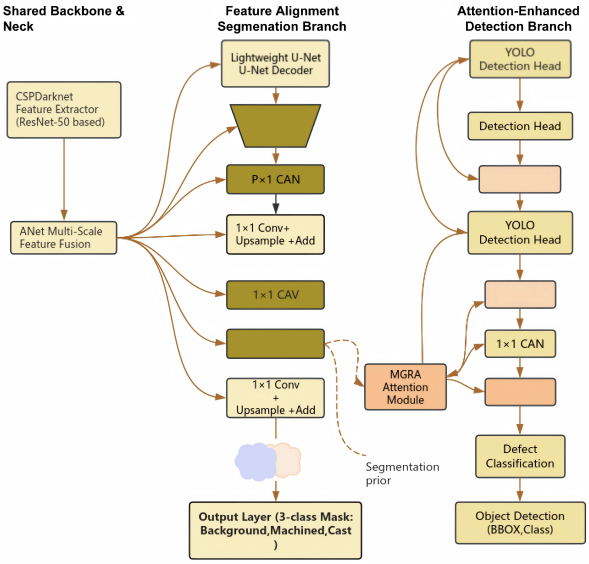


Figure 2: Architecture of the shared backbone and neck network.

3.2. *Shared backbone and neck*

An improved version of the cross-stage partial (CSP) Darknet, namely the C2f module, is adopted as the feature extractor. By introducing additional cross-layer connections, the C2f module enables richer gradient flow and enhances feature reuse. Combined with a

PANet-based multi-scale feature fusion strategy, the shared backbone is capable of capturing robust representations across different spatial resolutions. The parameters of this backbone are shared between the segmentation and detection tasks, allowing the network to learn task-agnostic visual representations while effectively reducing computational redundancy.The feature-aligned segmentation branch is implemented as a lightweight U-Net–style decoder, which maps multi-scale feature representations to pixel-wise region masks, including background, machining surfaces, and casting surfaces.The attention-enhanced detection branch incorporates a Mask-Guided Region Attention (MGRA) module before the conventional YOLO detection head. By leveraging the semantic priors provided by the segmentation branch, the MGRA module suppresses background noise and guides the detector to focus on critical machining regions. An overview of the proposed architecture is illustrated in Figure 2.

3.3. Segmentation Branch with Channel Alignment

Traditional YOLO models lack fine-grained pixel-level localization capabilities. To achieve surface segmentation without significantly increasing the computational burden, a lightweight segmentation decoder is designed, as shown in Figure 3. Unlike the standard

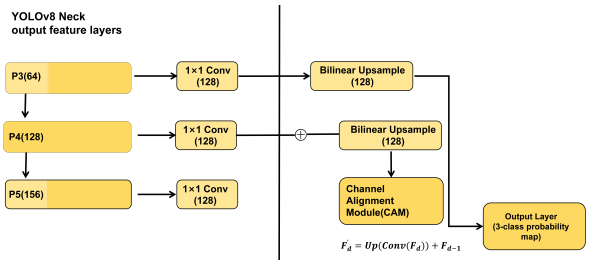


Figure 3: Structure of the segmentation module.

U-Net which directly concatenates features, the output feature layers (P3,P4,P5)) of the YOLOv8 neck have varying channel counts (64, 128, 256), leading to dimension mismatch during fusion. Therefore, we introduce the Channel Alignment Module (CAM) in the upsampling path. Specifically, for deep-layer features, a 1×1 convolution is first utilized to compress the channel dimensions to match the subsequent layer, followed by bilinear interpolation upsampling and element-wise addition:

$$F'_{i-1} = \text{Upsample}(\text{Conv}_{1\times1}(F_i)) + F_{i-1} \tag{1}$$

This design not only addresses the tensor dimension mismatch problem, but also acts as a bottleneck that effectively reduces the parameter complexity of the segmentation branch. The final fused features are then passed through the output layer to produce probability maps for the three region categories, as defined in Eq. (2).

$$M_{\text{pred}} \in \mathbb{R}^{H\times W\times C_{\text{seg}}} \tag{2}$$

3.4. Mask-Guided Residual Attention Module

The primary challenge in industrial defect detection is distinguishing actual flaws from texture noise on non-machined surfaces. To address this, we propose the MGRA module in Figure 4, which leverages segmentation semantic priors through a soft attention mechanism to maintain feature manifold continuity. The module comprises two parallel attention paths followed by residual fusion: Spatial Attention: To highlight func-

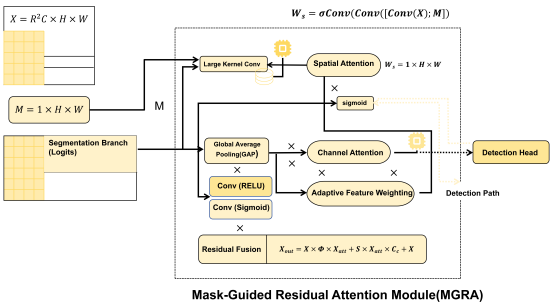


Figure 4: Structure of the MGRA module.

tional machined regions, the segmentation branch's logits (M_{pred}) serve as spatial priors. We apply large-kernel convolutions to capture local context, generating a spatial weight map $W_s \in \mathbb{R}^{H\times W\times 1}$:

$$W_c = \sigma\Big(\text{Conv}_{7\times7}\Big(M_{\text{pred}}\Big)\Big) \tag{3}$$

Channel Attention: To adaptively recalibrate defect-sensitive features, we apply Global Average Pooling (GAP) followed by a dual 1×1 convolutional structure. This produces channel weights $W_c \in \mathbb{R}^{1\times1\times C}$:

$$w_c = \sigma(\text{Conv}_2(\delta(\text{Conv}_1(\text{GAP}(F_{\text{in}}))))) \tag{4}$$

Residual Fusion:To prevent gradient vanishing-especially during early training when segmentation may be inaccurate-we employ a residual connection to fuse the weighted features with the original input F_{in}:

$$F_{\text{out}} = F_{\text{in}} + F_{\text{in}} \odot W_s \odot W_c \tag{5}$$

This dual-path mechanism forces the detection head to concentrate on segmented machined areas while preserving the robust representation capacity of the original backbone features.

3.5. Multi-Task Loss Function

To collaboratively optimize both tasks and address blurry edges in industrial parts, a composite loss function is introduced:

$$L_{\text{total}} = \lambda_{\text{seg}} L_{\text{seg}} + \lambda_{\text{bf}} L_{\text{bf}} + \lambda_{\text{det}} L_{\text{det}} \tag{6}$$

Segmentation Loss : A combination of Cross-Entropy and Dice Loss to balance pixel-level classification accuracy and regional overlap.

Boundary F1 Loss : To tackle the issue of non-sharp edges, we introduce a boundary loss using Sobel operators to compute gradient maps for both predicted and ground-truth masks, penalizing errors specifically in edge regions:

$$L_{\text{det}} = 1 - \frac{2 \sum \left(G_{\text{pred}} \cap G_{\text{gt}} \right)}{\sum G_{\text{pred}} + \sum G_{\text{gt}}} \tag{7}$$

Detection Loss : Includes CIoU Loss and DFL (Distribution Focal Loss) for bounding box regression, and VFL (Varifocal Loss) for classification.

3.6. Stability Strategy for Multi-Task Training

In multi-task learning, disparate convergence rates often trigger gradient dominance, potentially stalling the detection branch (0 mAP) during early joint optimization. To mitigate this, we propose Task-Specific Bias Initialization (TSBI). The classification bias b is calibrated to the defect prior probability π:

$$b = \log\left(\frac{\pi}{1 - \pi} \right) \tag{8}$$

By setting a low prior , the network is forced to output low initial confidence across all anchors. This prevents the loss explosion typically driven by an overwhelming majority of background samples. Consequently, TSBI restores gradient parity between tasks, facilitating stable and synchronized convergence.

4. Experiment preparation

4.1. Dataset and Experimental Setup

The proposed method was validated on an automated robotic inspection line in a real-world factory environment. The setup features a robotic arm that positions components before a high-magnification camera within a light-shielded enclosure to eliminate ambient interference. From 20,000 raw images collected on-site, a refined dataset of 9,000 samples was curated with dual annotations for detection and segmentation. The dataset covers five critical defect categories: padding damage, scratches, tool marks, rust, and slag holes. Standard preprocessing and data augmentation were applied to ensure model robustness and generalization.

4.2. Evaluation Metrics

To evaluate the multi-task performance of MGA-YOLO, we employ metrics focusing on segmentation accuracy, detection precision, and model complexity. Segmentation Metric: mIoU is used to quantify the geometric overlap between the predicted mask and the ground truth. Since the machined surface is the ROI , its segmentation accuracy directly dictates the effectiveness of subsequent detection:

$$\text{mIoU} = \frac{1}{K} \sum_{k=1}^{K} \text{IoU}_k \tag{9}$$

Detection Metrics: In addition to the standard mAP, we introduce Precision (P), Recall (R), and F1-score to gain deeper insights. In industrial inspection, Precision represents the proportion of true defects among all predicted samples, while Recall reflects the proportion of actual defects successfully captured by the model:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

We report mAP@50 and mAP@50:95 .The number of Parameters is used as the key indicator of spatial complexity to verify feasibility for edge device deployment.

4.3. Implementation Details

All models are trained with an input resolution of 640×640 for 300 epochs with a batch size of 32. We employ the SGD optimizer (momentum: 0.937, weight decay: 0.0005) and a cosine annealing schedule with an initial learning rate of 0.01. To mitigate gradient competition and imbalanced convergence inherent in multi-task learning, a Progressive Two-Stage Training Strategy is proposed:

Stage I: Encoder Warm-up. The detection head is frozen for the initial 20 epochs to prioritize the optimization of the backbone and segmentation branch. This prevents gradient oscillations caused by the high sensitivity of the untrained detection head to localization. L_{seg} and L_{bf} to guide the encoder in learning fundamental texture and edge features.

Stage II: Joint Optimization with Bias Correction. To address "task dominance" arising from the disproportionate magnitude of detection loss, we introduce Task-Specific Bias Initialization (TSBI). By re-initializing the classification layer's bias b, the initial detection loss is reduced by approximately two orders of magnitude, aligning the gradient scales of both tasks. Final joint fine-tuning is performed with loss weights $\lambda_{\text{det}} = 10.0$ and $\lambda_{\text{seg}} = 1.0$.

Table 1: Performance comparison of different methods on the test set.

Model	Precision	Recall	F1	mAP@0.50	mAP@0.5:0.95	Params (M)	FLOPs (G)	FPS
YOLOv8	81.4%	64.9%	72.2%	52.6%	38.5%	24.9	78.9	145
YOLOv5	73.2%	63.3%	67.9%	72.8%	42.4%	21.2	49.0	163
RT-DETR	82.5%	71.2%	76.4%	71.3%	45.5%	42.0	63.9	74
YOLOv11	76.1%	66.4%	70.9%	54.2%	38.6%	20.1	68.1	155
Faster R-CNN	85.8%	72.3%	78.1%	77.9%	49.8%	43.2	180.5	22
SSD	81.2%	69.7%	75.0%	76.8%	47.2%	26.3	123.7	45
Ours	**84.3**%	**77.2**%	**80.6**%	**76.7**%	**48.7**%	**25.86**	**79.1**	**135**

Table 2: Detection performance comparison on the test set.

Model	Precision	Recall	mAP@0.50	mAP@0.5:0.95	Params (M)	FLOPs (G)	FPS
Mask R-CNN	83.5%	73.4%	74.7%	46.1%	44.2	257.2	18
Mask DINO	86.3%	78.2%	80.5%	51.2%	52	286	23
YOLOv8m-Multi-Task	80.1%	71.5%	71.1%	39.8%	27.2	88.7	125
Ours	**84.3**%	**77.2**%	**76.7**%	**48.7**%	**25.86**	**79.1**	**135**

5. Experiment and Analysis

5.1. Comparative experiment

Detection Performance Analysis:MGA-YOLO was evaluated against state-of-the-art (SOTA) single-task detectors under identical hardware and labeling conditions. As shown in Table 1, MGA-YOLO matches or exceeds Faster R-CNN across all metrics while utilizing less than 60% of its parameters. Notably, it achieves a 5.5% improvement in mAP@50 and a 5.0% increase in mAP@50:95 over the single-task YOLOv8m baseline. This substantial gain is attributed to the Mask-Guided Attention (MGRA) mechanism, which effectively suppresses background-induced false positives in complex industrial scenes.

Multi-task Performance Comparison:Compared with heavy-duty multi-task networks like Mask R-CNN and Mask DINO Table 2, MGA-YOLO demonstrates significant architectural efficiency. Although Mask DINO provides high precision, its low FPS and heavy computational demand limit industrial deployment. Furthermore, compared to the detection-driven YOLOv8m-seg, which is prone to "segmentation collapse" upon detection failure, MGA-YOLO's balanced architecture ensures robust performance. Ultimately, MGA-YOLO achieves competitive accuracy with only 50% of the parameters and 28% of the FLOPs required by Mask DINO, proving its suitability for real-time edge-side applications.

Segmentation Performance Analysis: As summarized in Table 3, MGA-YOLO delivers superior segmentation results with an mIoU of 88.1% and a Dice score of 90.2%. Facilitated by feature alignment and the Boundary F1 loss, our method outperforms U-Net and DeepLabV3+ by 5.7% and 4.5% in mIoU, respectively, while maintaining a significantly lower parameter count than the latter. While BiSeNet v2 offers higher inference speeds, its accuracy fails to meet stringent industrial precision requirements. In contrast, MGA-YOLO strikes an optimal efficiency-accuracy trade-off, achieving 135 FPS with only 25.86M parameters.

Table 3: Performance comparison of the YOLOv11 series models

Model	**Type**	**mIoU**	**Dice**	**Params (M)**	**FPS**
U-Net	Seg-Only	82.4 %	88.1%	32.5	45
DeepLabV3+	Seg-Only	84.1%	89.5%	59.3	28
BiSeNet v2	Seg-Only	76.8%	83.2%	12.8	155
YOLOv8m-Seg	Multi-Task	81.5%	87.6%	24.9	125
Ours	**Multi-Task**	**88.1%**	**90.2%**	**25.86**	**135**

Table 4: Ablation study of the MGRA module.

Model	**Backbone**	**MGRA Module**	**P**	**R**	**mAP@0.50**
Baseline	YOLOv8m		81.4%	64.9%	71.2%
Ours	**YOLOv8m**	✓	**84.3**%	**77.2%**	**76.7%**

5.2. Ablation Study

Effectiveness of MGRA: Table 4 shows that incorporating the MGRA module yields a 5.5% mAP@50 boost. confirms it enhances target saliency and effectively suppresses false positives on complex casting backgrounds.

Training Strategy: Table 5 compares optimization schemes:

Case A (Naive): Without bias initialization, the model diverged early (0 mAP).

Table 5: Comparison of training strategies.

Case	Strategy	Init Bias	Loss Weights	Optimizer	Convergence	mAP@0.50
A	Naive Trainning		SGD	1.0/1.0	Diverged	0
B	Low Weight	✓	AdamW	1.0/0.05	Stable	35.6%
C	Ours	✓	Adam+SGD	0.5/0.15	Best	76.7%

Case B (Suboptimal): Despite convergence, improper weighting limited performance to 35.6% mAP.

Case C (Ours): By harmonizing gradients via TSBI and optimized weights, MGA-YOLO achieved a peak 76.7% mAP, proving gradient balancing is critical for multi-task stability.

5.3. Qualitative Analysis

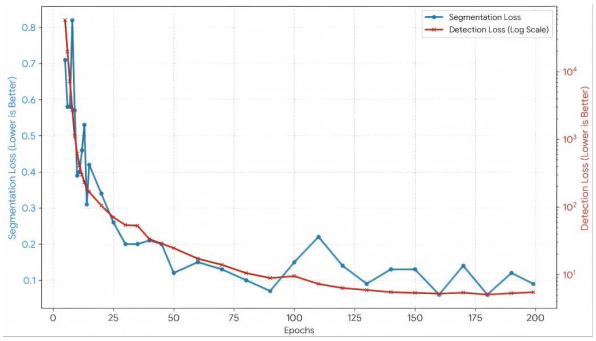
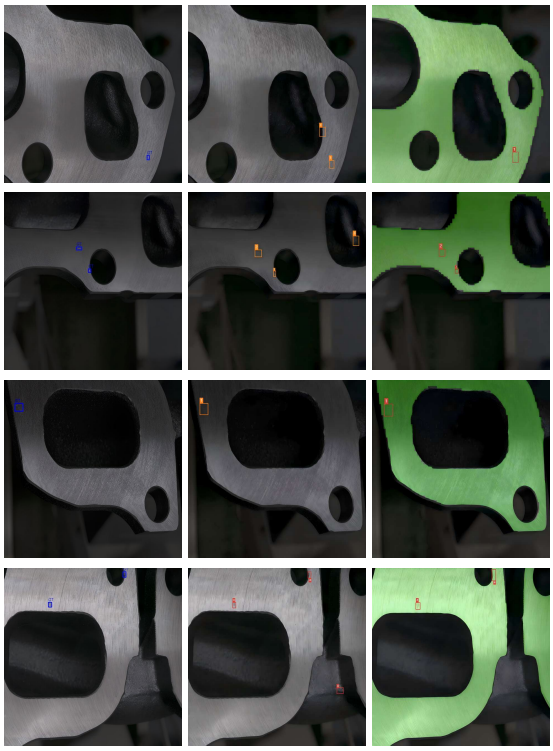


Figure 5: Curves of segmentation loss and detection loss during training.

Synergistic relationship between segmentation and detection: The loss trajectories in Figure 5 substantiate the claim that segmentation proficiency is a vital determinant of detection accuracy. Rather than being a secondary task, segmentation acts as the fundamental cornerstone of the proposed framework. The two curves exhibit a significant positive correlation in their convergence characteristics: significant optimizations in segmentation loss consistently trigger an accelerated convergence of the detection loss. This synchronization confirms that only by first establishing a robust representation of the surface geometry can the model achieve precise defect localization, validating the "segmentation-leads-detection" architectural logic.

In Figure 6 Qualitative comparison of (a) ground truth, (b) YOLOv8m, and (c) MGA-YOLO. While the YOLOv8m baseline suffers from small-object omissions and false positives on casting textures, MGA-YOLO achieves precise localization and high-fidelity segmentation. This superiority is primarily credited to the MGRA module, which leverages channel attention



(a) Input image (b) YOLOv8m (c) MGA-YOLO

Figure 6: Visualization results on the dataset: (a) Input images. (b) YOLOv8m predictions. (c) MGA-YOLO predictions.

to heighten sensitivity to subtle defect features. Furthermore, the Boundary F1 Loss ensures sharp delineation between machined and casting surfaces; this precise spatial constraint narrows the model's focus, effectively suppressing background interference even under non-uniform illumination.

6. Conclusion

This study presents MGA-YOLO, a lightweight multi-task framework that addresses background interference in industrial defect detection through MGRA and segmentation-prior guidance. Our model achieves a superior balance of accuracy and efficiency with 76.7% mAP@50 and 88.1% mIoU, validated against SOTA baselines. The introduction of TSBI stabilizes multi-task training by resolving gradient competition. Distinct from "black-box" detectors, MGA-YOLO offers enhanced interpretability by providing verifiable segmentation evidence. These features, combined with high real-time performance, make it a robust solution for edge-side deployment. Future work will focus on extending the framework to more diverse and complex multi-surface industrial environments.

Funding

This research did not receive any external funding and was performed as part of the authors' routine academic research.

References

[1] R. Bai, N. Jiang, L. Yu, et al., Research on industrial online detection based on machine vision measurement system, in: Journal of Physics: Conference Series, Vol. 2023, IOP Publishing, 2021, p. 012052.

[2] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, Vol. 28, 2015.

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: European conference on computer vision, Springer, 2016, pp. 21–37.

[4] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[5] S. U. Jia, J. I. A. Ze, Q. I. N. Yichang, et al., Improved yolov8 algorithm for industrial surface defect detection, Journal of Computer Engineering & Applications 60 (14) (2024).

[6] Y. Wang, J. Huang, M. S. K. Dipu, et al., Yolo-rlc: An advanced target-detection algorithm for surface defects of printed circuit boards based on yolov5, Computers, Materials & Continua 80 (3) (2024).

[7] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, in: Advances in neural information processing systems, 2017.

[8] A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[9] X. Chen, F. Wei, G. Zeng, et al., Conditional detr v2: Efficient detection transformer with box queries, arXiv preprint arXiv:2207.08914 (2022).

[10] Y. Zhao, W. Lv, S. Xu, et al., Detrs beat yolos on real-time object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 16965–16974.

[11] F. Li, H. Zhang, H. Xu, et al., Mask dino: Towards a unified transformer-based framework for object detection and segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 3041–3050.

[12] B. B. Gao, Learning to detect multi-class anomalies with just one normal image prompt, in: European Conference on Computer Vision, Springer, 2024, pp. 454–470.

[13] G. Gui, B. B. Gao, J. Liu, et al., Few-shot anomaly-driven generation for anomaly classification and segmentation, in: European Conference on Computer Vision, Springer, 2024, pp. 210–226.

[14] Y. Tu, B. Zhang, L. Liu, et al., Self-supervised feature adaptation for 3d industrial anomaly detection, in: European Conference on Computer Vision, Springer, 2024, pp. 75–91.

[15] J. Liu, G. Xie, J. Wang, et al., Deep industrial image anomaly detection: A survey, Machine Intelligence Research 21 (1) (2024) 104–135.

[16] J. Liu, Z. Yang, Y. Song, A two-stage anomaly detection framework: Towards low omission rate in industrial vision applications, Advanced Engineering Informatics 55 (2023) 101822.

[17] Q. Chen, H. g. Luo, H. Gao, et al., Progressive boundary guided anomaly synthesis for industrial anomaly detection, IEEE Transactions on Circuits and Systems for Video Technology (2024).

[18] X. Hu, X. Peng, D. Luo, et al., Diffumatting: Synthesizing arbitrary objects with matting-level annotation, in: European Conference on Computer Vision, Springer, 2024, pp. 396–413.

[19] K. Batzner, L. Heckler, R. König, Efficientad: Accurate visual anomaly detection at millisecond-level latencies, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2024, pp. 128–138.

[20] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[21] C. Kim, H. Shin, E. Hong, et al., Seg4diff: Unveiling open-vocabulary segmentation in text-to-image diffusion transformers, arXiv preprint arXiv:2509.18096 (2025).

[22] S. Huang, S. He, H. Qin, et al., Score: Scene context matters in open-vocabulary remote sensing instance segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 12559–12569.

[23] S. N. Rai, S. Karthik, M. I. Georgescu, et al., Road obstacle video segmentation, arXiv preprint arXiv:2509.13181 (2025).

[24] Y. Ye, Y. Wu, X. Luo, et al., Medseqft: Sequential fine-tuning foundation models for 3d medical image segmentation, arXiv preprint arXiv:2509.06096 (2025).

[25] Y. X. Chng, X. Qiu, Y. Han, et al., Semantic refocused tuning for open-vocabulary panoptic segmentation, arXiv preprint arXiv:2409.16278 3 (6) (2024).

[26] J. Noh, W. Jeon, J. Park, Dual interaction network with cross-image attention for medical image segmentation, Pattern Recognition Letters (2025).

[27] R. Cao, C. Song, B. Yang, et al., Adapting segment anything model for unseen object instance segmentation, arXiv preprint arXiv:2409.15481 (2024).

[28] S. Yang, H. Wang, Z. Xing, et al., Segdino: An efficient design for medical and natural image segmentation with dino-v3, arXiv preprint arXiv:2509.00833 (2025).

[29] Y. Li, Y. Wu, Y. Lai, et al., Meddinov3: How to adapt vision foundation models for medical image segmentation?, arXiv preprint arXiv:2509.02379 (2025).

[30] J. Wu, X. Li, H. Ding, et al., Betrayed by captions: Joint caption grounding and generation for open vocabulary instance segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 21938–21948.

[31] S. Jang, J. Yun, J. Kwon, et al., Dial: Dense image-text alignment for weakly supervised semantic segmentation, in: European Conference on Computer Vision, Springer, 2024, pp. 248–266.

[32] F. Yan, X. Jiang, Y. Lu, et al., Wavelet and prototype augmented query-based transformer for pixel-level surface defect detection, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 23860–23869.