2025 IEEE International Test Conference in Asia (ITC-Asia) | 979-8-3315-7193-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ITC-Asia67627.2025.00016

FALCO-WAFER: Feature-Aware Lightweight Contextual Detector for Wafer Defect Detection

Haotian Zhang[1*], Shurong Cao[1*], Ningmu Zou[1,2†],
[1]School of Integrated Circuits, Nanjing University, Suzhou, China
[2]Interdisciplinary Research Center for Future Intelligent Chips (Chip-X), Nanjing University, Suzhou, China

***Abstract*—Accurate wafer defect detection is critical for yield control in advanced semiconductor manufacturing. Traditional rule-based and CNN-based methods struggle with subtle, low-contrast anomalies, while transformer detectors often suffer from high computational overhead. We propose FALCO-WAFER, a lightweight, feature-aware detector tailored for region-of-interest patches from inline AOI systems. Our architecture combines a Multi-Scale Depthwise Block for efficient texture encoding with a Token-Energy Diagonal Attention head for robust feature refinement, enabling anchor-free inference at arbitrary resolutions. Evaluated on a real-world dataset with 5,723 labeled defect images, FALCO-WAFER achieves 90.7% AP@0.5 and 7.19% FNR using only 13.3M parameters, outperforming both CNN and transformer baselines. Its compact design supports low-latency deployment in high-throughput inspection lines. Code is available at: https://github.com/MrJoker06/FALCO-WAFER.**

***Keywords*—wafer inspection, AOI, lightweight detector, contextual modeling.**

I. Introduction

In semiconductor manufacturing, wafer defect detection plays a mission-critical role. Even minute structural anomalies, such as scratches, pattern dislocations, or particle contamination, can lead to the failure of entire dies or wafers, ultimately impacting product yield and system reliability. In high-dependability domains, such as automotive and aerospace electronics, undetected defects may propagate into functional faults, resulting in costly recalls and potential safety hazards. Therefore, it is imperative to perform accurate inspections before downstream processes, such as dicing and packaging, to ensure quality control and process stability.

Currently, wafer inspection systems typically rely on inline automatic optical inspection (AOI) tools to perform a coarse scan of the entire wafer and automatically crop multiple regions-of-interest (ROIs) that are suspected of containing defects [1]. These cropped image patches are then subjected to downstream classification or visual inspection. However, traditional rule-based approaches, such as statistical filters, defect pattern analysis, and process window monitoring, are constrained by fixed heuristics and limited granularity, making them poorly suited to detect subtle, irregular, or low-contrast defects. Even predictive screening methods that incorporate equipment logs or electrical test data suffer from challenges, including complex feature alignment, strong dependence on multi-sensor inputs, and limited scalability in practical environments.

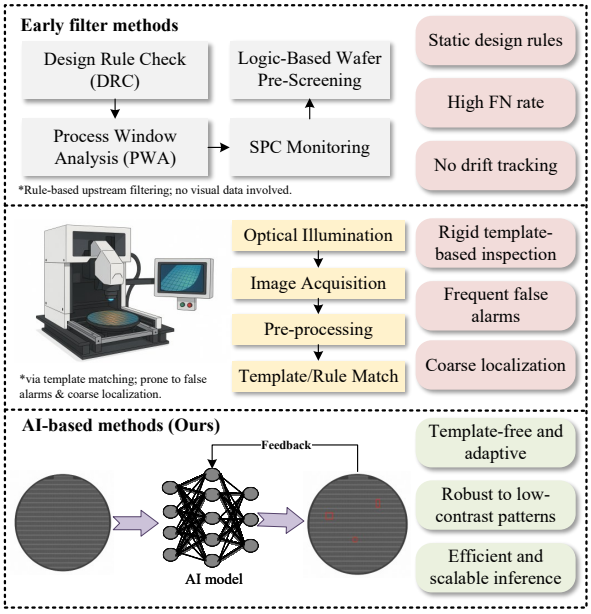

Fig. 1. Industrial Paradigms for Wafer Defect Inspection: From Early Screening to AI-Based Solutions.

In recent years, deep learning has emerged as a promising solution to address these limitations. CNN-based [2] and Transformer-based [3], [4] models have demonstrated significant improvements in identifying latent defects that traditional methods often miss. Nonetheless, most existing models still suffer from several drawbacks, such as heavy computation, dependence on anchor boxes or auxiliary classifiers, and slow inference speed, which hinder their real-time deployment in manufacturing lines.

To overcome the limitations of conventional detectors, we propose **FALCO-WAFER**, a feature-aware, lightweight object detection framework tailored for patch-level wafer inspection. Operating in a single-stage, anchor-free paradigm, FALCO-WAFER integrates a *Multi-Scale Depthwise Block* to capture anisotropic defect cues and a *Token-Energy Diagonal Attention* module to suppress low-saliency responses, enabling efficient and robust inference across varying input resolutions.

*These authors contribute equally to this work. †Corresponding author: nzou@nju.edu.cn. This work was supported in part by the National Natural Science Foundation of China (62341408).

979-8-3315-7193-1/25/$31.00 ©2025 IEEE
DOI 10.1109/ITC-Asia67627.2025.00016

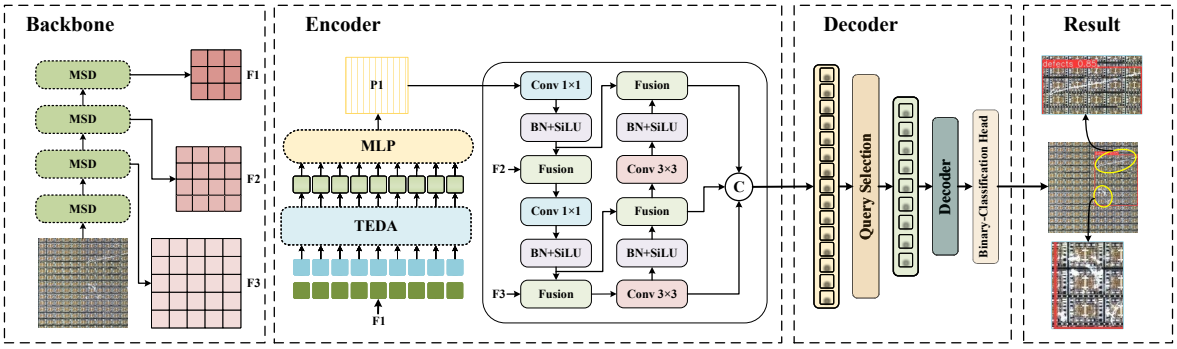

Fig. 2. The structure of the FALCO-WAFER framework.

The main contributions of this work are summarized as follows:

• **FALCO-WAFER Model.** We propose *FALCO-WAFER*, a lightweight, feature-aware, end-to-end detector designed for real-time wafer defect inspection near the end of manufacturing flow.

• **Dynamic feature extraction.** A multi-scale module is introduced to enhance fine-grained defect cues under noisy and low-contrast conditions, with minimal computational cost.

• **Efficient representation refinement.** We employ an expressiveness guided strategy to improve feature separability without introducing additional computation or complexity.

• **Industrial Validation.** We evaluate FALCO-WAFER on industrial wafer datasets, showing a superior trade-off between detection accuracy and inference speed, and confirming its practicality for near-line deployment.

II. Related Work

A. Wafer Defect Inspection in Industrial Practice

Within a fabrication line, optical inspection is the last gatekeeper before wafers enter back-end processing; microscopic surface defects, such as scratches, cracks, line-edge dislocations, particles, and edge chipping, must be screened to prevent downstream reliability crises. Conventional automated optical inspection (AOI) tools rely on edge detectors, template matching, and fixed-threshold filters. Although effective under stable illumination and layout conditions, their performance deteriorates when process drift, illumination variation, or new product designs alter the visual appearance of defects. Extensive rule-set tuning and per-product calibration further constrain scalability.

Recent deployments, therefore, favor data-driven inspection. Deep networks trained directly on wafer maps or ROI patches can learn discriminative patterns without relying on handcrafted rules, improving robustness across products and processes. As production takt times tighten, the demand grows for detectors that achieve high recall on subtle anomalies without incurring prohibitive inference costs.

B. AI-Driven Detection Frameworks for Wafer Defects

Hand-engineered algorithms built on geometric heuristics [5], such as edge continuity, gray-level statistics, or simple morphological rules, struggle with noise, pattern diversity, and scale variation. Supervised deep learning addresses these weaknesses: convolutional networks capture hierarchical texture cues; capsule networks [6] preserve part-whole relationships and offer limited interpretability. Yet, most early approaches treat detection as classification on sliding windows, which incurs redundant computation and latency.

Single-stage CNN detectors deliver faster inference but assume balanced object scales and structured layouts, conditions rarely met by sparsely distributed, low-contrast wafer defects. Their limited receptive field often misses faint or irregular anomalies. Transformer-based detectors remove anchors and model long-range context, but larger parameter footprints, slower convergence, and sensitivity to background noise hamper direct fab-floor deployment.

These constraints motivate the proposed **FALCO-WAFER** framework, which combines multi-scale depthwise feature extraction with a lightweight token-energy attention head, thereby eliminating anchors and multi-stage post-processing while retaining real-time throughput. The full architecture is detailed in Section III.

III. Proposed Methodology

A. Wafer-Aware Detection Framework

In wafer inspection, image data exhibit distinct structural properties: (1) a circular layout with radial illumination falloff near the edges, (2) a highly regular die lattice forming strong horizontal-vertical textures, and (3) sparse, small-scale defects often aligned along scribe lines or die borders.

These structural priors motivate **FALCO-WAFER**, a compact end-to-end detector tailored for AOI-cropped ROI patches. As shown in Fig. 2, it consists of:

- A convolutional backbone with **MSD** modules for efficient extraction of anisotropic textures.
- A **TEDA** module to enhance salient features and suppress noise.

- An anchor-free decoder for direct prediction of defects in a single pass.

B. Multi-Scale Depthwise Block (MSD)

To enhance the feature extraction capacity of the backbone while preserving computational efficiency, we introduce the **Multi-Scale Depthwise Block (MSD)**, a lightweight module that combines directional depthwise convolutions with adaptive dynamic mixing to construct expressive, multi-scale representations.

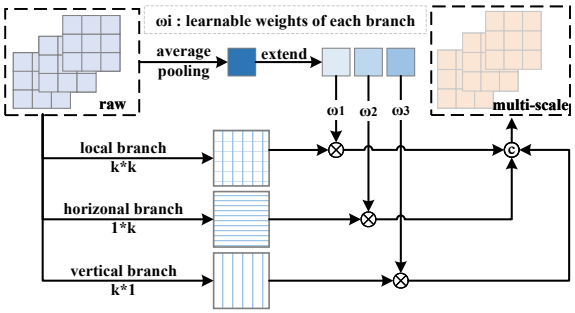

Fig. 3. Architecture of the Multi-Scale Depthwise Block.

Each MSD is constructed upon a novel *Dynamic Inception Mixer*, which processes input channels via a two-branch grouping strategy. The input tensor is first split along the channel dimension and then processed independently by a series of *Dynamic Inception Depthwise Convolutions*. Each convolutional path consists of three components: a square kernel ($k \times k$) for local context, and two directional band kernels ($1 \times k$ and $k \times 1$) to capture long-range dependencies along vertical and horizontal axes. These kernels operate in a depthwise fashion, significantly reducing parameter count and FLOPs.

To fuse the multi-path responses, we adopt a global context-aware weighting mechanism. Channel-wise weights are dynamically generated via a lightweight attention unit consisting of global average pooling followed by a 1×1 convolution and softmax activation. This produces adaptive coefficients that emphasize structurally salient patterns across spatial orientations:

$$\hat{x} = \sum_{i=1}^{3} \omega_i \cdot \mathrm{DWConv}_i(x), \tag{1}$$

$$\omega_i = \mathrm{Softmax}(W_{\mathrm{attn}}(\mathrm{GAP}(x))) \tag{2}$$

Empirically, MSD-blocks not only retain fine-grained local structures but also capture anisotropic context with high efficiency. This design is particularly effective in wafer-level defect inspection, where anomalies can appear at multiple scales and orientations under noisy conditions. By introducing strong inductive biases through directional depthwise convolutions and dynamic fusion, the MSD-block provides a powerful yet compact building block for scalable feature encoding in our overall architecture.

C. Token Energy-guided Diagonal Attention (TEDA)

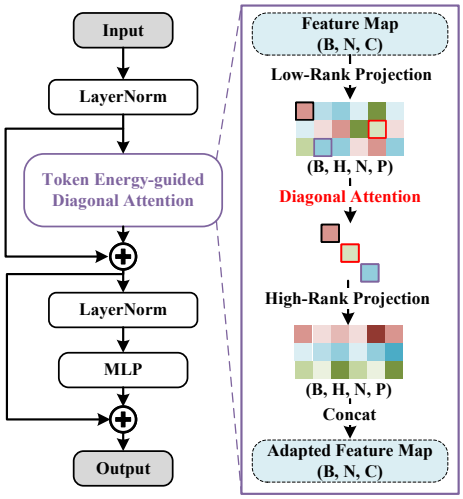

Fig. 4. Schematic of the TEDA Mechanism for Token-Level Feature Refinement.

In conventional vision transformers, the self-attention mechanism [7] typically computes token-to-token dependencies via dot-product similarity, which introduces quadratic computational complexity and tends to amplify irrelevant background noise in industrial imagery. To address these limitations and improve the flow of information within the feature encoding pipeline, we introduce the **Token Energy-guided Diagonal Attention (TEDA)** module. TEDA is conceptually motivated by the principle of maximizing mutual information (MI) between feature representations and task-relevant semantics [8], encouraging the network to retain only high-utility features during forward propagation.

TEDA replaces pair-wise attention with a *token-centric energy projection* strategy. Given feature tokens distributed over multiple attention heads, we first ℓ_2-normalize and square the vectors to form a second-order energy representation. For each head, a soft assignment mask $\Pi \in \mathbb{R}^{B\times H\times N}$ is generated by a temperature-scaled softmax on token energies, measuring the contribution of every token *inside that head*:

$$\Pi_{jk} = \mathrm{softmax}\Big(\textstyle\sum_d w_{jkd}^2/\tau\Big), \tag{3}$$

where τ is a learnable temperature and w_{jkd} denotes the d-th channel of token k in head j after normalization.

Instead of similarity-based attention, TEDA applies a self-suppressing diagonal weight:

$$\alpha_j = \frac{1}{1 + \sum_k \Pi_{jk}\,\|w_{jk}\|_2^2}, \tag{4}$$

and obtains the gated output per head by

$$z_{jk} = -\,\Pi_{jk}\,\alpha_j\,w_{jk}. \tag{5}$$

The head-wise outputs z_{jk} are finally **concatenated along the head dimension** and passed through a linear projection

TABLE I
DETECTION PERFORMANCE COMPARISON ACROSS MODELS

Model	GFLOPs(G)	Params (M)	mAP	AP@0.5	FNR (%)	FPS
Faster R-CNN	156.7	41.42	0.621	0.884	9.80	262
FCOS	112.0	19.10	0.591	0.858	31.20	273
YOLOv8-n	8.1	3.00	0.652	0.862	21.03	393
YOLO11-n	**6.5**	**2.60**	0.666	0.880	14.88	302
YOLOv8-s	28.5	11.13	0.659	0.874	15.02	387
YOLO11-s	21.5	9.40	0.671	0.880	20.17	292
DINO-r18	112.6	24.52	0.589	0.858	16.30	294
DINO-r50	208.1	47.54	0.629	0.853	14.40	267
Deformable-Detr-r18	33.6	22.26	0.328	0.585	14.11	286
Deformable-Detr-r50	145.0	40.10	0.571	0.845	**7.07**	286
Ours	47.4	13.30	**0.688**	**0.907**	7.19	210

to restore the original channel width, producing the updated token features.

This formulation offers three key advantages: (1) **Linear computational complexity** concerning sequence length, allowing scalable inference on high-resolution wafer images; (2) **Improved robustness to noise**, achieved by selectively suppressing low-energy tokens rather than performing exhaustive dense attention; (3) **Better semantic alignment** between intermediate features and detection targets, by focusing model capacity on informative spatial regions.

In practice, TEDA produces more compact and discriminative latent representations, particularly under challenging conditions such as low contrast or small defect size. These results suggest that TEDA not only reduces token interaction redundancy but also enhances mutual information between preserved features and ground-truth labels, an essential property for accurate and reliable real-time wafer defect inspection.

IV. EXPERIMENTS AND RESULTS

A. Dataset Description

We evaluate FALCO-WAFER on a custom dataset of wafer defects comprising more than **5,700** manually annotated images. The data are sourced from (i) *industrial wafers* captured on production lines, exhibiting authentic process noise and layout diversity, and (ii) *laboratory wafers* procured for in-lab analysis, each carrying naturally occurring defects [9].

A **7:1:2** split yields *4,006* training, *572* validation, and *1,145* test images, enabling the model to generalize across noisy industrial scenes and clean laboratory settings. **Due to inconsistent defect definitions across industrial settings, this work focuses on robust detection of potential defect regions rather than fine-grained classification. The proposed framework, however, remains compatible with multi-class tasks given standardized labels.** Moreover, for confidentiality reasons, commercial production-line images are not publicly visualized; only in-lab wafer samples are shown for illustration.

B. Model Comparison

All experiments are conducted on a workstation equipped with an NVIDIA RTX 4080 GPU (16GB). We benchmark FALCO-WAFER against representative CNN-based and transformer-based detectors, including Faster R-CNN, FCOS, YOLOv8, YOLO11, DINO, and Deformable-DETR variants. All models are trained under identical settings and evaluated using six metrics: GFLOPs, parameter count, mean Average Precision (mAP, reflecting localization and classification accuracy across IoU thresholds), AP@0.5 (accuracy at 0.5 IoU, commonly used in industrial quality control), false-negative rate (FNR, i.e., the proportion of actual defects missed by the detector), and throughput in FPS.

As shown in Table I, FALCO-WAFER achieves the best overall trade-off between accuracy and efficiency. It attains the highest AP@0.5 (**0.907**) and the lowest FNR (**7.19%**) with only 13.3M parameters—substantially outperforming both CNN and transformer baselines. In particular, compared to heavy-weight transformer models (e.g., DINO-r50 and Deformable-DETR-r50), our method yields a gain of over 4.5 percentage points in AP while reducing FLOPs and parameter count by more than 3×. Relative to lightweight YOLO-based models, FALCO-WAFER maintains comparable or lower computational complexity, but significantly reduces FNR by up to **14%**, indicating a lower miss rate for actual defect regions, critical for production reliability.

The reported FPS values only provide a reference for runtime efficiency; while absolute speed may vary across deployments, FALCO-WAFER's 210 FPS confirms its ability to support high-throughput, near real-time inspection under practical hardware constraints.

Notably, increasing backbone depth within the same model family (e.g., YOLOv8-n to YOLOv8-s, or DINO-r18 to DINO-r50) offers limited accuracy gains, underscoring that wafer defect detection may benefit more from localized texture awareness than deeper semantic abstraction in our case. These results affirm the value of our lightweight, wafer-aware design for industrial deployment scenarios.

C. Ablation Study

We conduct an ablation study to assess the individual contributions of the Multi-Scale Depthwise Block (MSD) and the Token Energy-Guided Diagonal Attention (TEDA). Starting from a base model rtdetr-r18 [10], introducing MSD significantly improves AP while reducing computational cost.

Adding TEDA further enhances detection accuracy and lowers the false negative rate. When combined, these two modules form the complete FALCO-WAFER architecture, achieving the highest overall performance.

TABLE II
RESULTS OF ABLATION STUDY

MSD	TEDA	GFLOPs	Params	AP@0.5	FNR
✗	✗	56.9	19.89	0.852	11.59
✓	✗	43.5	11.54	0.901	7.77
✗	✓	58.7	19.84	0.903	7.58
✓	✓	47.4	13.30	**0.907**	**7.19**

Fig. 5 shows representative wafer patches and corresponding activation maps under different module configurations. Compared to MSD-only and TEDA-only variants, the full FALCO model exhibits sharper and more localized responses around defect regions, demonstrating the complementary effect of multi-scale encoding and energy-based attention. Combined with its consistently superior detection performance over heavier detectors, these results validate FALCO-WAFER as a lightweight yet robust solution for high-throughput wafer-defect inspection.

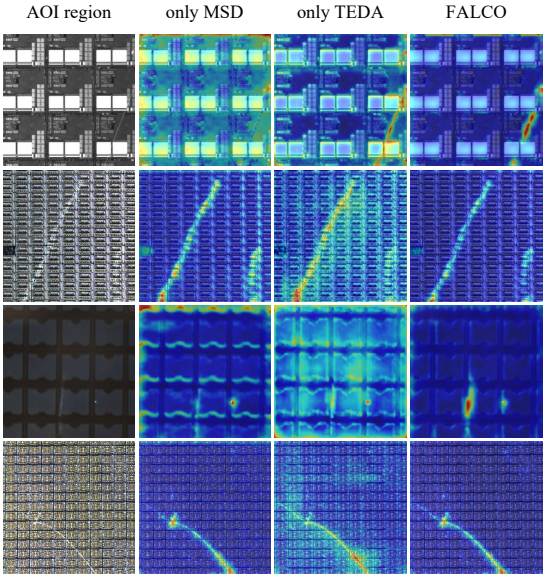

Fig. 5. Visual comparison of activation responses from different model variants.

D. Robustness to Realistic Disturbances

To evaluate visual robustness under production-like perturbations, we construct a synthetic *Drift-Set* by applying common image-level degradations to the test set. These include brightness and contrast shifts, mild Gaussian blur, and gamma adjustment, emulating practical inline variations such as illumination changes, lens wear, and sensor drift. All ground-truth annotations are retained to ensure consistency.

As shown in Table III, FALCO-WAFER exhibits strong tolerance to these disturbances. The AP@0.5 drops by only **1.6%**, and the FNR rises by less than **1%**, confirming the model's robustness under non-ideal conditions and its suitability for stable deployment without frequent AOI recalibration.

TABLE III
ROBUSTNESS ON CLEAN VERSUS DRIFTED IMAGES

Set	AP@0.5	FNR (%)
Clean-Set	0.907	7.19
Drift-Set	0.891	8.06

V. CONCLUSION

In this work, we presented FALCO-WAFER, a lightweight and high-precision object detector. We present FALCO-WAFER, a lightweight and context-aware detection framework tailored for real-time wafer defect inspection. Through multi-scale depthwise encoding and energy-guided attention, the model achieves robust performance under diverse visual conditions. Additional experiments on artificially perturbed inputs demonstrate strong tolerance to illumination and geometric shifts, though such distortions may not fully replicate the complexity of in-fab variations. As the current design assumes static spatial input and consistent labeling, future extensions may explore adaptation to temporal drift, cross-wafer dependencies, and evolving defect patterns in dynamic manufacturing environments.

REFERENCES

[1] F. Adly, P. D. Yoo, S. Muhaidat, Y. Al-Hammadi, U. Lee, and M. Ismail, "Randomized general regression network for identification of defect patterns in semiconductor wafer maps," *IEEE Transactions on Semiconductor Manufacturing*, vol. 28, no. 2, pp. 145–152, 2015.
[2] H. Park, J. Kim, and S. Choi, "Defectnet: Cnn-based surface defect detection for semiconductor wafers," *IEEE Trans. Semiconductor Manufacturing*, vol. 33, no. 2, pp. 246–254, 2020.
[3] S. Li, X. Lu, H. Zhang, and J. Sun, "Dino: Detr with improved denoising training," in *Proc. ICLR*, 2023.
[4] F. Mohammad and D. Ryu, "Semiconductor wafer map defect classification with tiny vision transformers," *arXiv preprint arXiv:2504.02494*, 2024.
[5] N. Yu, Q. Xu, and H. Wang, "Wafer defect pattern recognition and analysis based on convolutional neural network," *IEEE Transactions on Semiconductor Manufacturing*, vol. 32, no. 4, pp. 566–573, 2019.
[6] I. Fujita, Y. Nagamura, M. Arai, and S. Fukumoto, "Note on capsnet-based wafer map defect pattern classification," in *2021 IEEE 30th Asian Test Symposium (ATS)*, 2021, pp. 37–42.
[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[8] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *International conference on machine learning*. PMLR, 2018, pp. 531–540.
[9] T. Lei, B. Wang, S. Chen, S. Cao, and N. Zou, "Texture-ad: An anomaly detection dataset and benchmark for real algorithm development," 2024. [Online]. Available: https://arxiv.org/abs/2409.06367
[10] S. Li, X. Lu, H. Zhang, and J. Sun, "Rt-detr: Real-time detection transformers," *arXiv preprint arXiv:2304.08069*, 2023.