

Article

BMF-DETR: Pseudo-Depth-Guided Bidirectional Multi-Strategy Fusion for End-to-End Object Detection

Hai Wang ^{1,2} , Junhao Wen ^{1,2} , Chunlai Yang ^{1,2,*} , Kamara Kekele Adnan Fayçal ^{1,2} and Jiale Gu ^{1,2}

¹ School of Mechanical Engineering, Anhui Polytechnic University, Wuhu 241000, China; wanghai@ahpu.edu.cn (H.W.); wjh715wjh@outlook.com (J.W.); kekeleadnan@gmail.com (K.K.A.F.); 18862228705@163.com (J.G.)

² Anhui Key Laboratory of Advanced Numerical Control & Servo Technology, Wuhu 241000, China

* Correspondence: ycl@ahpu.edu.cn

Abstract

Transformer-based detectors model long-range context effectively, yet their representations remain dominated by RGB appearance and can become unreliable in cluttered, occluded, or crowded scenes. We present BMF-DETR, a pseudo-depth-guided detector that introduces RGB-derived geometric structure without requiring a depth sensor. DA3Mono-Large from Depth Anything 3 generates spatially aligned pseudo-depth maps offline, while two ResNet-50 streams encode appearance and relative geometry. Bidirectional cross-modal attention (BCMA) establishes two-way correspondence, and multi-strategy fusion (MSF) combines the streams through global calibration, channel allocation, and spatial gating before squeeze-and-excitation (SE) recalibration. On the fixed validation/evaluation split of the 2024 Roboflow-curated PASCAL VOC derivative, the complete model reaches 60.80 AP, compared with 52.80 AP for a capacity-matched dual-RGB control. BMF-DETR obtains 49.30 AP on COCO 2017. Its detector contains 58 M parameters and requires 103 GFLOPs; these figures exclude offline pseudo-depth generation. A shared-low-level variant retains 60.10 AP with 50 M parameters and 87 GFLOPs. The results show that pseudo-depth can serve as a useful auxiliary representation when its contribution is separated from capacity effects and evaluated under controlled fusion settings.

Keywords: end-to-end object detection; DETR; pseudo-depth; cross-representation fusion; bidirectional attention; geometric prior

1. Introduction

Building on the Transformer architecture [1], DETR reformulates object detection as a set-prediction problem and removes the need for hand-crafted anchors and non-maximum suppression [2]. Subsequent work has substantially improved its convergence and multi-scale representation. Deformable DETR limits attention to sparse sampling points [3]; Conditional DETR and DAB-DETR strengthen spatial priors in the query formulation [4,5]; DN-DETR and DINO improve optimization through denoising and better query initialization [6,7]; and RT-DETR demonstrates that an end-to-end detector can also be designed for real-time operation [8]. Despite these advances, the representation used by most DETR-family models remains largely appearance-driven.

Appearance cues are often sufficient in ordinary scenes, but they become less reliable when foreground and background share similar texture, when the visible part of an object is small, or when instances overlap heavily. Relative geometry offers a different description



Academic Editor: Olufemi A. Omitaomu

Received: 7 August 2026

Revised: 13 September 2026

Accepted: 17 September 2026

Published: 21 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

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

of the same scene: foreground–background ordering, contour continuity, and coarse spatial hierarchy. Sensor-based RGB-D systems exploit this information directly [9–13], but an additional depth sensor increases cost, synchronization requirements, and deployment complexity. Modern monocular depth models, including MiDaS, Depth Anything, and Depth Anything V2 [14–16], provide another option by deriving a depth-like representation from an RGB image alone.

Pseudo-depth, however, is not an independent physical observation. It is inferred from the same RGB image and therefore remains correlated with appearance. It may also contain local errors around small objects, weak boundaries, reflective surfaces, and severe occlusion. The key question is consequently not whether pseudo-depth can simply be added to an RGB detector, but whether the detector can extract useful structure from this derived representation while limiting duplicated or unreliable responses. This distinction matters because otherwise gains can be wrongly attributed to geometry when they may partly come from an extra backbone, additional parameters, or an easier optimization path.

BMF-DETR addresses this issue by retaining separate RGB and pseudo-depth streams long enough to establish bidirectional correspondence, followed by content-dependent fusion across global, channel, and spatial dimensions. The contribution lies in this organization of heterogeneous-but-correlated representations; cross-attention, SE recalibration, and monocular depth estimation themselves are established techniques. The experimental design therefore includes capacity matching, shuffled and degraded pseudo-depth, simple fusion, unidirectional interaction, MSF component ablations, fusion-level ablations, runtime accounting, lightweight variants, and depth-model adaptation.

The main contributions are threefold. First, we formulate a two-stage pseudo-depth-assisted detection pipeline in which DA3Mono-Large generates reusable depth maps offline and the detector learns from paired RGB and cached pseudo-depth inputs. Second, we organize bidirectional residual interaction, global calibration, channel allocation, spatial gating, and channel recalibration across three semantic levels. Third, controlled experiments separate the effects of auxiliary content, network capacity, fusion direction, fusion strategy, and detector cost. The contribution is this tested organization of correlated representations rather than a claim of a new attention operator.

2. Related Work

2.1. DETR-Based End-to-End Object Detection

DETR combines a Transformer encoder–decoder, learnable object queries, and Hungarian matching to predict a set of object instances directly [2]. The original formulation is conceptually simple but converges slowly and is weak on small objects. Deformable DETR addresses both issues by replacing dense global attention with sparse, reference-point-based sampling [3]. Conditional DETR and DAB-DETR alter query representation to improve spatial conditioning [4,5], whereas DN-DETR and DINO stabilize matching and accelerate optimization through denoising [6,7]. These methods provide the detector framework used in this study but continue to build their feature hierarchy mainly from RGB appearance.

2.2. Pseudo-Depth and RGB-Derived Auxiliary Representations

Depth information has long been used to complement RGB in detection and segmentation [9–13]. Sensor depth and infrared images are independent observations with their own noise characteristics. Pseudo-depth is different: it is generated from RGB and therefore inherits both the visual evidence and the priors learned by the monocular depth model. MiDaS improves cross-dataset relative-depth generalization [14], while Depth Anything and Depth Anything V2 scale this idea using large amounts of unlabeled imagery [15,16].

This makes pseudo-depth attractive when no depth sensor or manual depth annotation is available, but it also means that any gain must be interpreted as better use of a derived representation rather than as the addition of a new physical modality.

Recent object-detection studies have begun to exploit this setting directly. Li et al. [17] generate pseudo-depth with Depth Anything V2 and use a dual-branch detector, while Hou and Liu [18] introduce depth-guided quality cues into an aerial detector. These works are closer to BMF-DETR than conventional RGB-D methods. Our difference lies in the organization of interaction and fusion: the two streams are updated bidirectionally at three semantic levels and are subsequently fused through global calibration, channel allocation, spatial gating, and SE recalibration before Deformable-DETR-style set prediction.

2.3. Heterogeneous Representation Interaction and Fusion

Concatenation, element-wise addition, and fixed-weight fusion are easy to implement but cannot adapt the relative contribution of appearance and geometry to object scale, occlusion, or scene content. Attention-based modules such as SENet and CBAM show that feature channels and spatial positions do not contribute equally [19,20]. Cross-attention further allows one representation to query another, which has been exploited in RGB-X segmentation, multispectral detection, and token-level multimodal fusion [12,13,21–23].

The challenge in pseudo-depth fusion is more specific because the two streams share the same visual source. Their relationship is simultaneously complementary and redundant. BMF-DETR does not explicitly estimate depth uncertainty, and its learned weights should not be read as calibrated reliability scores. Instead, the architecture provides a flexible interaction and allocation mechanism whose behavior is evaluated empirically using capacity, shuffling, degradation, and simple-fusion controls.

3. Materials and Methods

3.1. Overall Architecture

BMF-DETR follows a two-stage pipeline. In the preparation stage, DA3Mono-Large from Depth Anything 3 [24,25] is run once for each RGB image and the resulting pseudo-depth map is cached. The model uses a ViT-L encoder and a DPT prediction head; the weight identifier is depth-anything/DA3MONO-LARGE. During detector training and evaluation, paired RGB and cached pseudo-depth inputs pass through two ResNet-50 streams, three interaction-fusion levels, multi-scale projection, and a Deformable-DETR-style encoder–decoder. The depth generator is outside the detector optimization graph in the main configuration.

$$D = \Phi_d(I) \quad (1)$$

Figure 1 shows the complete data flow. The selected fusion locations are denoted L1-L3 throughout the manuscript and correspond to the second, third, and fourth output stages of the two ResNet-50 branches, with 512, 1024, and 2048 channels, respectively. At each level, BCMA first updates the streams in both directions; MSF then performs global, channel, and spatial allocation, and SE recalibrates the fused output before its projection to the detector.

$$\hat{F}_i = \text{SE}(\text{MSF}(\text{BCMA}(F_i^{\text{rgb}}, F_i^d))), i \in \{1, 2, 3\} \quad (2)$$

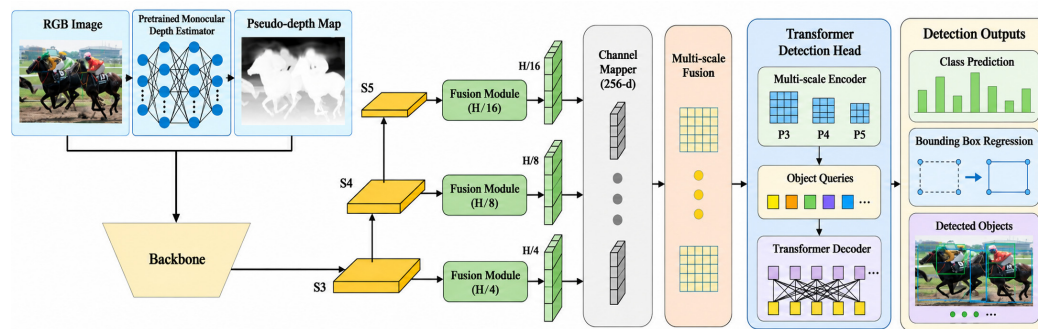


Figure 1. Overall architecture and two-stage data flow of BMF-DETR. Pseudo-depth is prepared offline and cached before detector training. Colors distinguish schematic components and do not encode quantitative measurements.

3.2. Offline Pseudo-Depth Construction and Dual-Stream Encoding

Pseudo-depth is produced by single-image DA3Mono-Large inference with $\text{process_res} = 756$. The prediction is restored to the original image size by bilinear interpolation. For each image, valid responses are clipped at the 2nd and 98th percentiles, linearly mapped to $[0, 255]$, and stored as an 8-bit grayscale PNG, with nearer regions represented by larger values. The detector loader applies a second per-image 1st/99th-percentile clipping step to the cached PNG, maps the result to $[0, 255]$, repeats it over three channels, and standardizes it with mean 127.5 and standard deviation 127.5. These maps encode relative scene structure rather than metric depth.

$$D_c = \text{clip}(D, q_{0.02}, q_{0.98}), D_g = 255 \frac{D_c - D_{\min}}{D_{\max} - D_{\min} + \epsilon}, \epsilon = 1 \times 10^{-6} \quad (3)$$

RGB images, cached pseudo-depth maps, and annotations are paired by image identifier. The RGB and depth arrays are concatenated into a six-channel sample before augmentation, so the same random flip, resize, crop, and bounding-box transformation is applied to both representations. Training uses a horizontal flip probability of 0.5 and multi-scale resize/crop branches; no RGB-only color jitter is used. This shared geometric path preserves pixel-level alignment between the two inputs and the boxes. Figure 2 shows examples of paired RGB images and pseudo-depth maps.



Figure 2. Examples of RGB images and their corresponding pseudo-depth representations.

Both branches use independently parameterized ResNet-50 backbones initialized from ImageNet [26,27]. The first stage and its normalization statistics are frozen. Fusion is applied to the next three outputs, whose channel widths are 512, 1024, and 2048. Replicating

the grayscale pseudo-depth map over three channels adds no information; it only preserves compatibility with the pretrained backbone interface.

$$F_i^{\text{rgb}} = E_i^{\text{rgb}}(F_{i-1}^{\text{rgb}}), F_i^d = E_i^d(F_{i-1}^d) \quad (4)$$

3.3. Bidirectional Cross-Modal Attention

As illustrated in Figure 3, BCMA establishes two-way correspondence before irreversible fusion. Each feature map is adaptively pooled to 14×14 , flattened into tokens, and layer-normalized. Eight attention heads are used at every fusion level. In one direction, RGB tokens query pseudo-depth keys and values; in the other direction, pseudo-depth tokens query RGB keys and values. The two attended outputs are restored to the original feature resolution and added residually to their respective branches.

$$X_i^{\text{rgb}} = \text{LN}(\text{Flatten}(\text{Pool}(F_i^{\text{rgb}}))), X_i^d = \text{LN}(\text{Flatten}(\text{Pool}(F_i^d))) \quad (5)$$

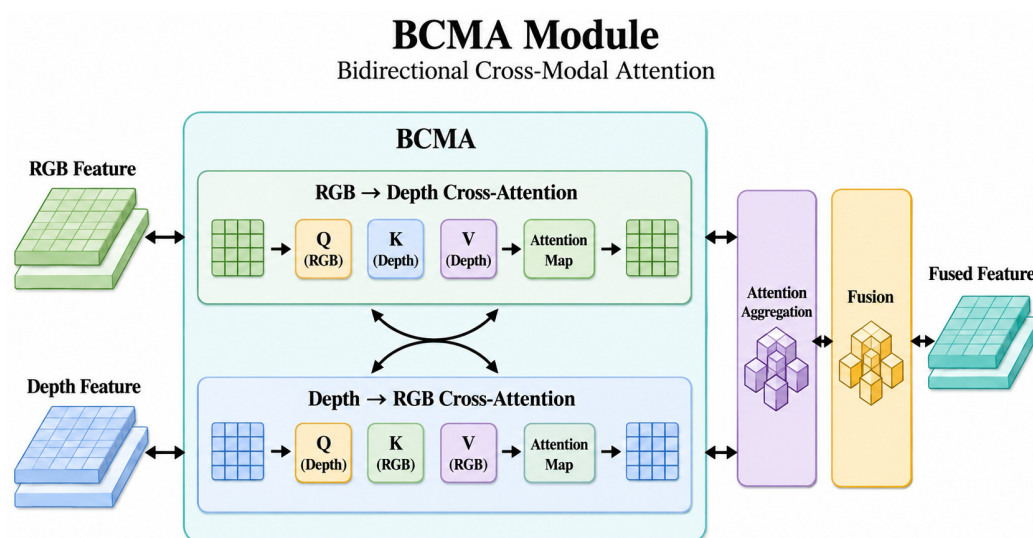


Figure 3. Structure of the bidirectional cross-modal attention (BCMA) module.

In the depth-to-RGB path, RGB tokens query pseudo-depth keys and values; in the reverse path, pseudo-depth tokens query RGB keys and values. The attended tokens are reshaped to spatial maps, bilinearly restored to the original feature resolution, and added residually to the corresponding branch. This residual form preserves branch-specific information and delays irreversible fusion until MSF.

$$Y^{\text{rgb} \leftarrow d} = \text{MCA}(X^{\text{rgb}}, X^d, X^d), Y^{d \leftarrow \text{rgb}} = \text{MCA}(X^d, X^{\text{rgb}}, X^{\text{rgb}}) \quad (6)$$

$$\tilde{F}_i^{\text{rgb}} = F_i^{\text{rgb}} + \text{Up}(\text{Reshape}(Y^{\text{rgb} \leftarrow d})), \tilde{F}_i^d = F_i^d + \text{Up}(\text{Reshape}(Y^{d \leftarrow \text{rgb}})) \quad (7)$$

3.4. Multi-Strategy Fusion and Channel Recalibration

BCMA outputs two enhanced but still separate streams. As shown in Figure 4, MSF converts them into a unified representation in three steps. First, global descriptors from one stream calibrate the other, introducing sample-level cross-branch conditioning. Second, global and local descriptors are used to allocate channel contributions competitively between RGB and pseudo-depth. Third, a spatial gate allocates the two representations at

each location. The local pooling kernel is $k = 3$, and GroupNorm uses 32 groups whenever the channel count permits.

$$z_i^{\text{rgb}} = \text{GAP}(\tilde{F}_i^{\text{rgb}}), z_i^{\text{d}} = \text{GAP}(\tilde{F}_i^{\text{d}}) \quad (8)$$

$$[\omega_i^{\text{rgb}}, \omega_i^{\text{d}}] = \text{Softmax}(\Psi([s_i^{\text{global}}, s_i^{\text{local}}])), \omega_i^{\text{rgb}} + \omega_i^{\text{d}} = 1 \quad (9)$$

MSF Module

Multi-Strategy Fusion

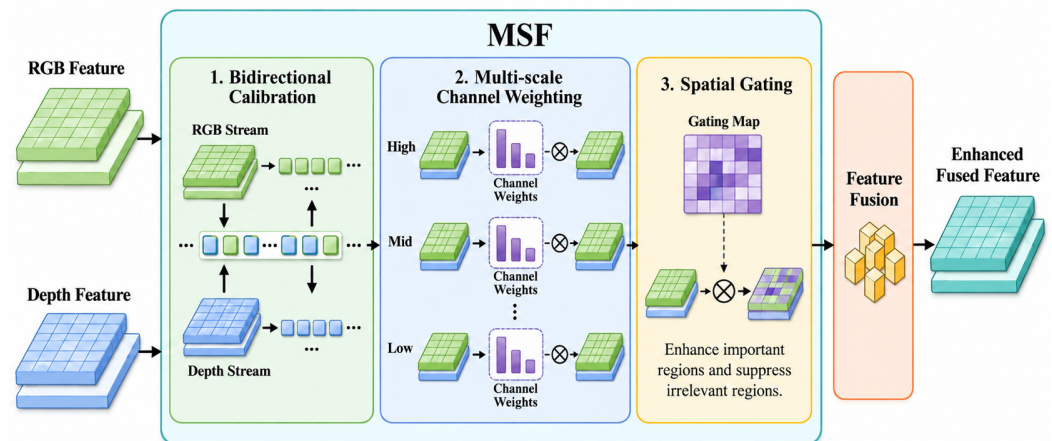


Figure 4. Structure of the multi-strategy fusion (MSF) module.

The weighted streams are concatenated, normalized, and projected by a 1×1 convolution. A standard SE block with reduction ratio 16 then recalibrates the fused channels [19]. The weights are learned through the detection objective only; they are not explicit estimates of depth confidence. Figure 5 summarizes the complete BCMA–MSF–SE fusion module.

$$F_i^{\text{fuse}} = \text{Conv}_{1 \times 1}(\text{GN}(\text{Concat}(F_i^{\text{rgb}*}, F_i^{\text{d}*}))), \hat{F}_i = \text{SE}(F_i^{\text{fuse}}) \quad (10)$$

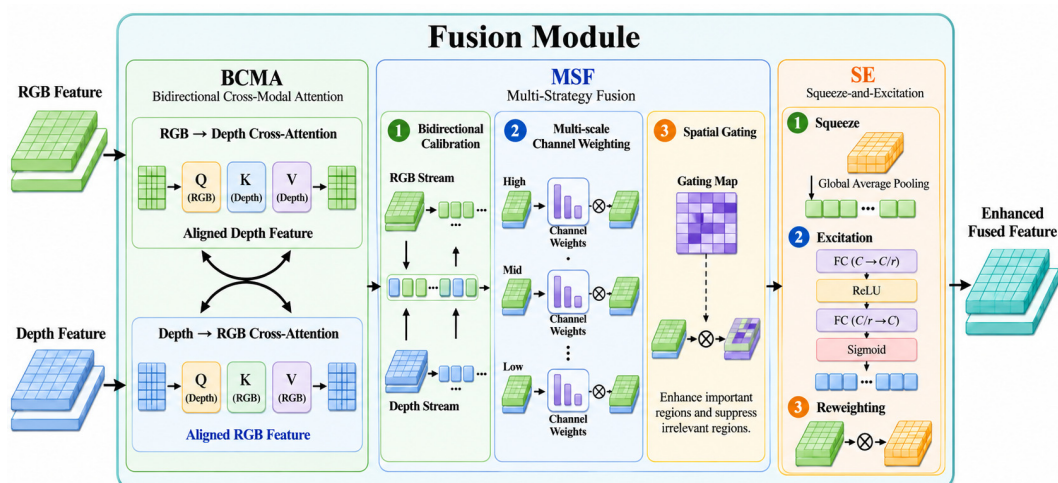


Figure 5. Organization of BCMA, MSF, and SE within a fusion block.

3.5. Multi-Scale Set Prediction and Optimization

The three fused outputs are projected to 256 channels. A fourth, lower-resolution feature level is generated by a stride-2 projection from the highest-level fused feature. The Deformable-DETR encoder and decoder each contain six layers, and the decoder

uses 300 object queries. These settings are kept fixed for the RGB baseline and controlled ablations.

Training uses Hungarian bipartite matching. The classification term is Focal Loss, while box regression combines L1 and generalized IoU losses. The loss weights are $\lambda_{\text{cls}} = 2$, $\lambda_{\text{L1}} = 5$, and $\lambda_{\text{GIoU}} = 2$; the same coefficients are used in the matching cost. Auxiliary supervision is applied to intermediate decoder layers.

$$L = 2L_{\text{cls}} + 5L_{\text{L1}} + 2L_{\text{GIoU}} \quad (11)$$

Given the paired RGB and cached pseudo-depth tensors, the detector performs set prediction without anchors or non-maximum suppression. The auxiliary adaptation study in Section 4.5 changes the depth-training boundary and is reported separately from the main offline configuration.

4. Results and Discussion

4.1. Experimental Settings and Reproducibility

Experiments use a Roboflow-curated PASCAL VOC 2012 public derivative exported on 6 February 2024 [28], rather than the official VOC 2012 train/validation/test partition [29,30]. Its COCO-format annotations contain 17,112 images from the 2007–2012 VOC releases: 13,690 training images with 31,356 boxes and 3422 validation images with 7835 boxes. RGB, pseudo-depth, and annotation file counts agree one-to-one, and no overlap was found between the two splits after the removal of the Roboflow export suffix from the original image identifiers. The validation split is evaluated after every epoch, is used to select the checkpoint with the highest coco/bbox_mAP, and supplies the final reported VOC-derived results; no independent test set is used. COCO experiments use train2017 and val2017. Evaluation follows the COCO bounding-box protocol [31,32]: AP averages IoU thresholds from 0.50 to 0.95 in steps of 0.05, and AP50, AP75, APs, APm, and APl are reported separately.

All detector variants are trained for 50 epochs with AdamW [33], an initial learning rate of 2×10^{-4} , weight decay of 1×10^{-4} , and gradient clipping at 0.1. The first 500 iterations use linear warm-up from a factor of 0.001; the learning rate is multiplied by 0.1 at epochs 40 and 48. The training batch size is 3 per training process, the validation batch size is 2, and automatic learning-rate scaling is disabled. Both backbones use ImageNet initialization, while the detector is initialized from a COCO-pretrained two-stage Deformable DETR checkpoint.

Training samples use shared RGB-depth resizing with short-side choices of 480, 512, 544, 576, 608, and 640 pixels and a maximum long side of 800 pixels. The alternate augmentation branch resizes to 640, 704, or 768 pixels, applies a 384×600 random crop, and resizes again. Validation uses a ratio-preserving 800×640 resize. Depth preparation retains its separate process_res = 756 setting, which should not be confused with the detector input policy. Table 1 summarizes the implementation settings.

For the cross-paper comparison, DETR, Deformable DETR, DN-DETR, DAB-DETR, and REGO-DETR retain values reported in their respective publications, whereas CF-DETR is a 50-epoch local reproduction. The supplementary comparison protocol records the result source, backbone, input convention, feature levels, training duration, and augmentation provenance. Because these settings are heterogeneous, the cross-paper comparison in Section 4.3 is used only for context. Claims about the fusion design rely on the controlled experiments in Section 4.2.

Table 1. Experimental environment and key settings.

Item	Configuration
Workstation	Dell workstation
OS	Ubuntu 24.04
CPU	Intel Core i7-12400F
GPU	NVIDIA A6000
Memory	128 GB
Storage	Samsung 970 EVO Plus 3 TB
Framework	PyTorch 1.12.1; CUDA 11.6
Detector input	training short side 480–640; maximum long side 800; validation 800 × 640
Depth generator	Depth Anything 3 DA3Mono-Large; process_res = 756; offline cache
VOC-derived dataset	Roboflow PASCAL VOC 2012 public derivative; export dated 6 February 2024; 17,112 images
Data split	train 13,690; validation/evaluation 3422; no separate independent test set
Optimization	AdamW; 50 epochs; batch = 3 per training process; no automatic LR scaling

The gray header row identifies column labels; shading has no quantitative meaning.

4.2. Controlled Ablation Studies

We begin with the original module-combination sequence. The complete model reaches 60.80 AP compared with 49.10 AP for the RGB baseline. This 11.70-point difference is not assigned to BCMA alone because the pseudo-depth branch and additional capacity enter the model together. Table 2 therefore serves as a component-combination record, while Tables 3 and 4 isolate capacity, representation content, fusion strategy, MSF operations, and fusion placement.

Table 2. Core module combinations on the fixed validation/evaluation split of the Roboflow-curated VOC-derived dataset.

RGB	BCMA	MSF	SE	AP	AP50	AP75	APs	APm	API	ΔAP
✓				49.10	68.70	53.30	13.70	31.60	61.10	—
✓	✓			57.60	73.50	63.20	15.40	38.30	70.80	+8.50
✓	✓		✓	58.70	74.80	64.10	16.90	40.20	71.30	+9.60
✓	✓	✓		59.90	76.60	65.30	15.60	42.60	72.80	+10.80
✓	✓	✓	✓	60.80	76.60	65.70	17.90	42.50	72.90	+11.70

The gray header row identifies column labels; shading has no quantitative meaning.

The final SE step is not uniformly beneficial across every metric. AP rises from 59.90 to 60.80 and APs from 15.60 to 17.90, but APm changes from 42.60 to 42.50 and AP50 remains 76.60. We therefore describe SE as a modest refinement rather than as a universal gain across scales.

The capacity-matched control is more informative. When the second branch receives another copy of RGB and all remaining architecture is unchanged, AP is 52.80, well below the complete 60.80 AP. Shuffling pseudo-depth across images reduces AP to 50.70, while controlled pseudo-depth degradation yields 57.10. Simple concatenation and element-wise addition reach 54.60 and 53.80 AP, and a CBAM-based gating baseline reaches 56.60. These results indicate that neither a second backbone nor generic fusion is sufficient to reproduce the complete-model gain.

Table 3. Controlled representation and fusion ablations under a common training schedule.

Variant	AP	AP50	AP75	APs	APm	API
RGB only	49.10	68.70	53.30	13.70	31.60	61.10
Pseudo-depth only	43.20	62.10	46.00	10.80	27.90	55.40
Concat + 1×1	54.60	71.80	58.90	14.60	35.70	66.20
Element-wise add	53.80	71.10	57.80	14.20	35.10	65.50
No BCMA (MSF + SE)	56.20	73.30	60.70	15.40	37.60	68.30
Depth→RGB only	58.90	75.40	63.70	16.20	40.00	71.40
RGB→Depth only	58.40	75.00	63.10	16.00	39.50	70.90
Dual-RGB capacity control	52.80	70.20	56.70	14.10	34.10	64.20
Shuffled pseudo-depth	50.70	69.30	54.60	13.50	32.90	62.30
Degraded pseudo-depth	57.10	73.90	61.70	15.10	38.70	69.60
MSF only	55.10	72.50	59.50	14.90	36.40	67.10
SE only	50.30	69.40	54.20	14.20	32.20	61.90
CBAM gating baseline	56.60	73.60	61.10	15.20	37.90	68.80
BMF-DETR (full)	60.80	76.60	65.70	17.90	42.50	72.90

The gray header row identifies column labels; shading has no quantitative meaning.

Table 4. MSF-operation and fusion-level ablations.

MSF Variant	AP	Fusion Levels	AP
Full MSF	60.80	L1 only	57.40
w/o global calibration	59.60	L2 only	58.10
w/o channel allocation	59.20	L3 only	57.70
w/o spatial gating	58.90	L1 + L2	59.10
—	—	L1 + L3	58.80
—	—	L2 + L3	59.40
—	—	L1 + L2 + L3	60.80

The gray header row identifies column labels; shading has no quantitative meaning.

Directionality also matters: depth→RGB and RGB→depth interactions reach 58.90 and 58.40 AP, respectively, both below the bidirectional model. Removing global calibration, channel allocation, or spatial gating from MSF reduces AP by 1.20, 1.60, and 1.90 points. No single fusion level matches the three-level design; L2 + L3 is the strongest two-level configuration at 59.40 AP.

4.3. Comparison with End-to-End Detectors on COCO 2017

Table 5 provides a contextual comparison on COCO 2017. BMF-DETR reaches 49.30 AP, 68.40 AP50, and 52.90 AP75. Its 58 M parameters and 103 GFLOPs cover the detector with prepared RGB and pseudo-depth tensors and exclude DA3Mono-Large. CF-DETR is the only external baseline reproduced locally for 50 epochs; the remaining rows retain their published values. Differences across these heterogeneous rows are therefore not treated as controlled estimates of the fusion gain.

Detailed provenance and protocol descriptors for every contextual comparison row are provided in Appendix A.1. Published rows retain their source authors' training and augmentation settings, whereas CF-DETR is identified as our local reproduction. Its parameter count and GFLOPs remain unreported because they were not recomputed with the same complexity script. The causal interpretation of BMF-DETR therefore rests on the common-code-path ablations in Section 4.2.

Table 5. Contextual comparison with representative end-to-end detectors on COCO 2017.

Method	Epochs	GFLOPs	Params (M)	AP	AP50	AP75	APs	APm	API
DETR [2]	500	86	41	42.0	62.4	44.2	20.5	45.8	61.1
Deformable DETR [3]	50	173	40	43.8	62.6	47.7	26.4	47.1	58.0
DN-DETR [6]	50	94	44	44.1	64.4	46.7	22.9	48.0	63.4
DAB-DETR [5]	50	216	44	45.7	66.2	49.0	26.1	49.4	63.1
CF-DETR [34] (reproduced)	50	96	45	48.2	67.1	52.7	31.5	51.0	63.2
REGO-DETR [35]	50	190	54	47.6	66.8	51.6	29.6	50.6	62.3
BMF-DETR (ours)	50	103	58	49.3	68.4	52.9	30.3	52.3	66.2

The gray header row identifies column labels; shading has no quantitative meaning.

Among the listed results, BMF-DETR has the highest AP, AP50, AP75, APm, and API, while the reproduced CF-DETR result has slightly higher APs. Because the rows differ in implementation and input policy, this pattern is reported descriptively and is not used to infer a specific advantage for small or large objects.

4.4. Computational Cost and Reporting Boundary

The reported 58 M parameters and 103 GFLOPs describe the BMF-DETR detector with cached pseudo-depth already available. This boundary includes the two ResNet-50 streams, BCMA, MSF, SE, multi-scale projection, encoder, decoder, and detection heads. It does not include DA3Mono-Large preprocessing, depth inference, resizing, or cache generation.

Table 6 retains only measurements that can be matched to the detector boundary. Under the recorded A6000 setting, the RGB baseline detector takes 24.8 ms per image, and the cached-input BMF detector takes 35.7 ms per image. DA3Mono-Large and the complete RGB-to-depth-to-detection path were not remeasured under the same tool, precision, and input protocol; their parameters, FLOPs, latency, throughput, and peak memory are therefore not reported. The manuscript makes no real-time or complete-pipeline efficiency claim. A protocol with 100 warm-up iterations and 1000 timed iterations is suggested for future reproduction, but those counts are not presented as an experiment performed here or as a journal requirement.

Table 6. Detector complexity and runtime under the cached-input boundary.

Scope	Params	GFLOPs	Latency	Eq. FPS	Peak VRAM
RGB baseline detector	40 M	73	24.8 ms	40.3	5.6 GB
BMF detector (cached depth)	58 M	103	35.7 ms	28.0	7.8 GB

The gray header row identifies column labels; shading has no quantitative meaning. Eq. FPS is calculated as 1000 divided by latency in milliseconds and is not a separately measured throughput.

4.5. Lightweight Design and Depth-Model Adaptation

Three detector-side variants are evaluated to address the cost of the dual-stream design. C1 halves the width of the pseudo-depth branch after the stem; C2 retains only the L2 and L3 fusion blocks; and C3 shares the stem and first residual stage between the two streams while keeping higher-level encoders separate. All complexity values use the same detector-only boundary as the complete model. Detailed values are reported in Appendix A.2.

C1 provides the lowest detector cost, whereas C3 gives the strongest accuracy-efficiency compromise: it removes 8 M parameters and 16 GFLOPs while losing only 0.70 AP. We therefore regard C3 as the preferred deployment-oriented variant rather than claiming that the full model is intrinsically lightweight.

We also examine whether the pseudo-depth representation can be adapted to the detection task. The main configuration keeps DA3Mono-Large frozen and uses a fixed offline cache. D1 trains a lightweight residual adapter on the cached depth response. D2 is an isolated online-training variant that partially unfreezes the depth model; it is not used to explain the behavior of the main offline model or to support deployment claims. The corresponding results are reported in Appendix A.3.

The adaptation gains are modest and are not presented as statistically significant because repeated-run variance is not available. Their purpose is to show that task-aware depth updating is feasible. The frozen offline configuration remains the main model because it decouples depth preparation from detector training and allows the same auxiliary representation to be reused in all controlled ablations.

5. Conclusions

This study examined how an RGB-derived pseudo-depth representation can be incorporated into a DETR-style detector without treating it as an independent sensor modality. BMF-DETR keeps appearance and relative geometry in separate streams, establishes bidirectional correspondence at three feature levels, and then performs global, channel, and spatial allocation before channel recalibration. The method is therefore best understood as an organization of correlated representations rather than as a new attention primitive.

The controlled experiments clarify the source of the gain. The capacity-matched dual-RGB control reaches 52.80 AP, shuffled pseudo-depth 50.70 AP, concatenation 54.60 AP, and element-wise addition 53.80 AP, compared with 60.80 AP for the complete model. Unidirectional interaction is also weaker than bidirectional interaction. These results show that both the content of the paired pseudo-depth representation and the way it is integrated matter for performance.

On COCO 2017, BMF-DETR reaches 49.30 AP. The cached-input detector requires 58 M parameters and 103 GFLOPs, excluding DA3Mono-Large. A shared-low-level variant reduces the detector to 50 M parameters and 87 GFLOPs while retaining 60.10 AP on the Roboflow-derived validation/evaluation split. The depth-adaptation variants reach 61.10 and 61.30 AP, but they are reported as auxiliary single-run studies rather than statistically established improvements.

The main limitations are the absence of an independent VOC-derived test split, the reuse of the validation set for checkpoint selection and final reporting, the lack of repeated-run uncertainty estimates, and the missing unified cost profile for the complete DA3Mono-Large pipeline. Pseudo-depth may also be unreliable for very small, heavily occluded, reflective, or weak-texture objects. Future work should evaluate a held-out test protocol, quantify run-to-run variance, measure the complete pipeline under a fixed timing protocol, and investigate explicit uncertainty modeling and parameter sharing.

Author Contributions: Conceptualization, H.W. and C.Y.; methodology, H.W. and J.W.; software and validation, H.W. and J.W.; investigation, H.W., K.K.A.F. and J.G.; writing—original draft preparation, H.W.; writing—review and editing, all authors; supervision, C.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Overseas Study Visit and Training Program for Outstanding Young Backbone Talents of Anhui Province, grant number gxgwfx2021035; the Graduate Student Innovation Project of Anhui Province, grant number 2022xscx097; the Anhui Polytechnic University–Jiujiang District Industrial Collaborative Innovation Special Fund Project, grant number 2022cyxtb4; the Science and Technology Project of Wuhu City, grant number 2023yf092; and the Enterprise Cooperation Project of Anhui Future Technology Research Institute, grant number 2023qyhz17.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: PASCAL VOC and COCO 2017 are publicly available from their respective repositories. The VOC-based experiments use the Roboflow PASCAL VOC 2012 public derivative exported on 6 February 2024 [28]. The exact 13,690/3422 annotations, pseudo-depth file lists, configuration files, and evaluation scripts used in this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to acknowledge the financial support from the Major Science and Technology Special Projects in Anhui Province. The authors also acknowledge the use of OpenAI Codex (desktop version 26.915.4065.0; OpenAI, San Francisco, CA, USA) for assistance with manuscript formatting and language presentation.

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

Appendix A. Supplementary Protocol and Extension Results

Appendix A.1. Provenance of Contextual Comparison Results

Table A1 records the source and principal protocol descriptors of each row in the contextual COCO comparison. The heterogeneous settings are documented for transparency and are not interpreted as a controlled estimate of the fusion gain.

Table A1. Provenance and protocol descriptors for the COCO comparison rows.

Method	Result Source	Backbone/Variant	Input/Feature Levels	Epochs	Augmentation
DETR	Published [2]	ResNet-50	short side 800, max 1333; 1 level	500	source default
Deformable DETR	Published [3]	ResNet-50	800/1333; 4 levels	50	source default
DN-DETR	Published [6]	ResNet-50	800/1333; 4 levels	50	source default
DAB-DETR	Published [5]	R50-DC5, 3 patterns	800/1333; source feature config.	50	source default
CF-DETR	Re-trained by us [34]	ResNet-50	800/1333; 4 levels	50	our fixed pipeline
REGO-DETR	Published [35]	ResNet-50	800/1333; 4 levels	50	source default
BMF-DETR	This work	dual ResNet-50	short side 480–640, max 800; 4 detector levels	50	shared RGB-depth geometric augmentation

The gray header row identifies column labels; shading has no quantitative meaning.

Appendix A.2. Lightweight Detector Variants

Table A2 provides the detailed accuracy and detector-only complexity values for the three lightweight variants discussed in Section 4.5.

Table A2. Accuracy–complexity trade-off of lightweight variants.

Model	AP	Params	GFLOPs
BMF-DETR	60.80	58 M	103
C1: narrow depth branch	59.90	47 M	82
C2: fewer fusion levels	59.70	52 M	90
C3: shared low-level encoder	60.10	50 M	87

The gray header row identifies column labels; shading has no quantitative meaning.

Appendix A.3. Depth-Representation Adaptation

Table A3 reports the isolated depth-adaptation variants. These single-run results are supplementary and are not used to support the main frozen-offline configuration.

Table A3. Depth-representation adaptation study.

Configuration	Depth Update	AP	Δ AP
Frozen offline DA3Mono-Large (main)	No	60.80	—
Frozen depth + residual adapter	Adapter only	61.10	+0.30
Partial depth fine-tuning (isolated variant)	Final DPT block + head	61.30	+0.50

The gray header row identifies column labels; shading has no quantitative meaning.

References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
2. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *ECCV*; Springer: Cham, Switzerland, 2020; pp. 213–229.
3. 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 ICLR, Virtual*, 3–7 May 2021.
4. Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; Wang, J. Conditional DETR for Fast Training Convergence. In *Proceedings of the ICCV, Virtual*, 11–17 October 2021; pp. 3651–3660.
5. Liu, S.; Li, F.; Zhang, H.; Yang, Y.; Li, S.; Zhu, J.; Zhang, L. DAB-DETR: Dynamic Anchor Boxes Are Better Queries for DETR. In *Proceedings of the ICLR, Virtual*, 25–29 April 2022.
6. Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L.M.; Zhang, L. DN-DETR: Accelerate DETR Training by Introducing Query DeNoising. In *Proceedings of the CVPR, New Orleans, LA, USA*, 19–24 June 2022; pp. 13619–13627.
7. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.-Y. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. In *Proceedings of the ICLR, Kigali, Rwanda*, 1–5 May 2023.
8. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Y.; Liu, Q.; Chen, J. DETRs Beat YOLOs on Real-Time Object Detection. In *Proceedings of the CVPR, Seattle, WA, USA*, 17–21 June 2024; pp. 16965–16974.
9. Gupta, S.; Girshick, R.; Arbeláez, P.; Malik, J. Learning Rich Features from RGB-D Images for Object Detection and Segmentation. In *Proceedings of the ECCV, Zurich, Switzerland*, 6–12 September 2014; pp. 345–360.
10. Hu, X.; Yang, K.; Fei, L.; Wang, K. ACNet: Attention Based Network to Exploit Complementary Features for RGBD Semantic Segmentation. In *Proceedings of the ICIP, Taipei, Taiwan*, 22–25 September 2019; pp. 1440–1444.
11. Chen, X.; Lin, K.-Y.; Wang, J.; Wu, W.; Qian, C.; Li, H.; Zeng, G. Bi-directional Cross-Modality Feature Propagation with Separation-and-Aggregation Gate for RGB-D Semantic Segmentation. In *Proceedings of the ECCV, Glasgow, UK*, 23–28 August 2020; pp. 561–577.
12. Zhang, J.; Liu, H.; Yang, K.; Hu, X.; Liu, R.; Stiefelhagen, R. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation with Transformers. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 14679–14694. [[CrossRef](#)]
13. Shen, J.; Chen, Y.; Liu, Y.; Zuo, X.; Fan, H.; Yang, W. ICAFusion: Iterative Cross-Attention Guided Feature Fusion for Multispectral Object Detection. *Pattern Recognit.* **2024**, *145*, 109913. [[CrossRef](#)]
14. Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; Koltun, V. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 1623–1637. [[CrossRef](#)] [[PubMed](#)]
15. Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; Zhao, H. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *Proceedings of the CVPR, Seattle, WA, USA*, 17–21 June 2024; pp. 10371–10381.
16. Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; Zhao, H. Depth Anything V2. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 21875–21911. [[CrossRef](#)]
17. Li, S.-Q.; Feng, W.; Liu, B. Pseudo-depth-based deep neural network model for object detection. *Sci. Rep.* **2026**, *16*, 15120. [[CrossRef](#)] [[PubMed](#)]
18. Hou, X.; Liu, W. DGQ-YOLO: Depth-Guided Quality-Aware detection with Pseudo-Depth for aerial imagery. *J. King Saud Univ. Comput. Inf. Sci.* **2026**, *38*, 774. [[CrossRef](#)]
19. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the CVPR, Salt Lake City, UT, USA*, 18–22 June 2018; pp. 7132–7141.
20. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Proceedings of the ECCV, Munich, Germany*, 8–14 September 2018; pp. 3–19.
21. Wang, Y.; Chen, X.; Cao, L.; Huang, W.; Sun, F.; Wang, Y. Multimodal Token Fusion for Vision Transformers. In *Proceedings of the CVPR, New Orleans, LA, USA*, 19–24 June 2022; pp. 12186–12195.
22. Xiao, Y.; Meng, F.; Wu, Q.; Xu, L.; He, M.; Li, H. GM-DETR: Generalized Multispectral Detection Transformer with Efficient Fusion Encoder for Visible-Infrared Detection. In *Proceedings of the CVPRW, Seattle, WA, USA*, 17–18 June 2024; pp. 5541–5549.

23. An, R.; Guo, Y.; Wang, Z.; Zhao, Z.; Deng, C.; Li, J. MRT-DETR: A Robust Visible-Infrared Object Detector with Adaptive Cross-Modal Feature Fusion. *Infrared Phys. Technol.* **2026**, *153*, 106346. [CrossRef]
24. ByteDance-Seed. Depth Anything 3: Official Project and API. Available online: <https://github.com/ByteDance-Seed/Depth-Anything-3> (accessed on 13 September 2026).
25. Lin, H.; Chen, S.; Liew, J.H.; Chen, D.Y.; Li, Z.; Shi, G.; Feng, J.; Kang, B. Depth Anything 3: Recovering the Visual Space from Any Views. *arXiv* **2025**, arXiv:2511.10647.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the CVPR, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
27. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the CVPR, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
28. Roboflow. PASCAL VOC 2012 Object Detection Dataset, Version 1, COCO Export, 6 February 2024. Available online: <https://public.roboflow.com/object-detection/pascal-voc-2012> (accessed on 23 April 2026).
29. Everingham, M.; Eslami, S.M.A.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The PASCAL Visual Object Classes Challenge: A Retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]
30. PASCAL VOC Project. The PASCAL Visual Object Classes Challenge 2012. Available online: <https://www.robots.ox.ac.uk/~vgg/projects/pascal/VOC/voc2012/> (accessed on 11 September 2026).
31. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the ECCV, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
32. COCO Consortium. COCO API Bounding-Box Evaluation Implementation. Available online: <https://github.com/cocodataset/cocoapi/blob/master/PythonAPI/pycocotools/cocoeval.py> (accessed on 11 September 2026).
33. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the ICLR, New Orleans, LA, USA, 6–9 May 2019.
34. Cao, X.; Yuan, P.; Feng, B.; Niu, K. CF-DETR: Coarse-to-Fine Transformers for End-to-End Object Detection. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 185–193. [CrossRef]
35. Chen, Z.; Zhang, J.; Tao, D. Recurrent Glimpse-Based Decoder for Detection with Transformer. In Proceedings of the CVPR, New Orleans, LA, USA, 19–24 June 2022; pp. 5260–5269.

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.