DAMCSeg: Dynamic Adaptive Multi-Modal Collaborative Semantic Segmentation

Qirui Liao[1], Zuohua Ding[2], Hongyun Huang[3]*
School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310000, Zhejiang, China[1,2]
Library, Center of Multimedia Data Analysis, Zhejiang Sci-Tech University, Hangzhou 310000, Zhejiang, China[3]

***Abstract*—While current semantic segmentation models excel in controlled environments, they often struggle with key challenges such as dynamic multi-modal data, small target recognition, and computational efficiency for edge deployment. Motivated by these limitations, this study explores targeted solutions and presents DAMCSeg (Dynamic Adaptive Multi-modal Collaborative Semantic Segmentation), an innovative framework that introduces advancements across feature fusion, training paradigms, and model efficiency. The core contributions of DAMCSeg include: 1) a Dual-Stage Attention Fusion (DSAF) module that dynamically adjusts multi-branch fusion weights based on scene complexity; 2) an end-to-end joint training framework for object detection and semantic segmentation designed to minimize inter-stage error propagation; and 3) a Lightweight Multi-Modal Fusion (LMMF) module that efficiently integrates multi-source data with low computational overhead. To rigorously evaluate the proposed method's effectiveness against these specific challenges, extensive experiments are conducted on mainstream benchmark datasets. The results demonstrate that DAMCSeg achieves high accuracy and operational efficiency, effectively addressing critical issues in dynamic scene adaptation, complex target segmentation, and edge device deployment. This provides a practical and viable solution for semantic segmentation in demanding applications such as autonomous driving and medical image analysis.**

Keywords*—Dynamic Adaptive Multi-Modal Collaborative; Dual-Stage Attention Fusion (DSAF); End-to-End Detection-Segmentation Joint Framework; Lightweight Multi-Modal Fusion (LMMF)

I. Introduction

As a fundamental pillar of computer vision, semantic segmentation performs pixel-wise classification of images, enabling machines to interpret visual scenes with fine-grained understanding. This technology has become indispensable in real-world applications including autonomous driving, medical imaging, and intelligent surveillance systems. While deep learning—particularly through CNN and Transformer architectures—has dramatically advanced the state of the art, current segmentation models still encounter several critical limitations. For instance, the pursuit of high-level semantic representation often comes at the cost of weakened spatial detail preservation, which hampers accuracy on small objects and intricate boundaries. Moreover, conventional multi-branch fusion approaches frequently rely on fixed weighting schemes, lacking the flexibility to dynamically adapt to varying scene conditions. The dependence on single-modal data further restricts robustness under challenging environments like low illumination or significant occlusion. Additionally, many high-accuracy models incur substantial computational costs, hindering their deployment in real-time or resource-constrained settings.

Current mainstream semantic segmentation models can be divided into three categories: first, classic CNN-based models (e.g., DeepLab v3+ [1], PSPNet[2]), which improve context modeling capabilities through modules such as dilated convolution and pyramid pooling, but have shortcomings in small target segmentation and dynamic scene adaptation; second, Transformer-based models (e.g., SegFormer [3], Mask2Former [4]), which enhance feature correlation through global attention mechanisms, but suffer from slow inference speed and high computational cost; third, lightweight models (e.g., ENet [5], MobileNet-based [6] segmentation networks), which meet real-time requirements but have room for improvement in segmentation accuracy and robustness. Although recent multi-modal semantic segmentation approaches aim to enhance robustness under sensor degradation, many still assume complete modality inputs or employ static fusion strategies. For instance, early methods like FusionNet [7] focus on LiDAR completion rather than adaptive segmentation, while more recent works such as MM-Seg [8] and AdaFuse [9] explicitly address missing-modality scenarios through cross-modal prompting and learnable gating, respectively. However, these models operate with uniform fusion policies across the entire image and do not leverage cross-task geometric priors (e.g., from object detection) to resolve ambiguities in occluded or small-scale regions—a critical limitation in autonomous driving scenarios. This gap motivates our design of a dual-stage, detection-aware fusion mechanism.

To address the above issues, this paper proposes the DAMCSeg method. While components like attention-based fusion, joint training, and multi-modal integration exist in prior work, our key novelty lies in their synergistic combination and the specific design choices that enable a fundamental leap in adaptive capability, rather than mere incremental improvement. Our core innovations are as follows:

- A Dual-Stage Attention Fusion (DSAF) module is introduced, which uniquely decouples adaptation into two complementary levels: global scene-level assessment and local target-level refinement. In contrast to existing single-stage or scene-only aware fusion approaches (e.g., CFNet [10])—which impose a uniform fusion strategy across the entire image—DSAF enables simultaneous handling of simple backgrounds and highly complex, occluded objects within the same frame. This two-tiered mechanism directly addresses the challenge of decoupled global-local complexity, a gap unmet by current frameworks.

*Corresponding author.

- An end-to-end detection-segmentation joint training framework is developed, featuring a novel Detection-Segmentation Alignment Loss. This loss function explicitly penalizes geometric misalignment between predicted bounding boxes and segmentation masks, thereby mitigating error propagation inherent in cascaded or loosely coupled systems. The framework thus establishes a transferable principle for enhancing cross-task consistency through direct optimization of spatial alignment.
- A Lightweight Multi-Modal Fusion (LMMF) module is proposed, capable of fusing RGB with depth or infrared data while incorporating a modal missing complementation mechanism. This design ensures robust inference under partial sensor failure—a critical requirement for real-world deployment that most existing multi-modal models overlook due to their reliance on complete input modalities.
- Targeted loss functions, including a tiny target-focused loss and an occluded region-aware loss, are formulated to provide explicit supervisory signals for the most challenging segmentation scenarios. These losses significantly enhance model robustness in extreme conditions involving small-scale or heavily occluded objects.

Experimental results show that DAMCSeg significantly outperforms current mainstream semantic segmentation models on multiple benchmark datasets. Compared with DeepLab v3+ [1], the average mIoU is increased by 7.0% and the inference speed by 25.7%; compared with SegFormer-B4 [3], the mIoU is increased by 6.6% and the small target mIoU by 14.8%; compared with Mask2Former [4], the inference speed is increased by 128.9% while the mIoU is increased by 3.4%; in extreme scenarios such as low light, the mIoU is increased by 5%-8% compared with single-modal mainstream models. This method breaks the balance bottleneck among accuracy, speed, and robustness of existing mainstream models, providing a new idea for the practical application of semantic segmentation technology.

The main contributions of this paper are summarized as follows:

- Propose a Dynamic Adaptive Multi-Modal Collaborative Semantic Segmentation method (DAMCSeg), which realizes dynamic optimization of multi-branch fusion through a dual-stage attention mechanism to adapt to different scenes and target characteristics;
- Construct an end-to-end joint training framework for object detection and semantic segmentation, and design a cross-task loss function to improve the segmentation accuracy of complex targets and boundary regions;
- Develop a lightweight multi-modal fusion module, supporting multi-source data integration and model compression to achieve coordinated optimization of high accuracy and real-time performance;
- Conduct extensive experiments on multiple benchmark datasets, comparing comprehensively with current mainstream semantic segmentation models to verify the superiority of the proposed method, and provide a practical solution for semantic segmentation in complex and extreme scenarios.

II. Related Work

A. Object Detection

Object detection technology is an important support for semantic segmentation, and its development has evolved from traditional handcrafted feature-based methods to deep learning-based methods. Traditional methods such as the Viola-Jones detector and HOG + SVM rely on manually designed features and have limited generalization capabilities. In the deep learning era, two mainstream paradigms have been formed: two-stage detectors represented by the R-CNN [11] series, which achieve high-precision detection through “region proposal + fine-grained classification”; and single-stage detectors represented by YOLO [12] and SSD [13], which pursue inference speed through “end-to-end direct prediction”. In recent years, DETR [14] has introduced the Transformer architecture to realize set prediction without NMS post-processing, and models such as Faster R-CNN [15] enhanced with FPN structures have improved multi-scale detection capabilities.

Some existing mainstream semantic segmentation models (e.g., Mask2Former [4]) integrate object detection ideas, but fail to achieve deep collaborative training between detection and segmentation, resulting in mismatches between localization accuracy and semantic segmentation accuracy. Drawing on the advantages of end-to-end joint training, this paper deeply integrates object detection and semantic segmentation, realizing mutual promotion through a shared feature backbone and cross-task loss function, and providing accurate localization priors for complex target segmentation.

B. Semantic Segmentation

The field of semantic segmentation has formed diverse technical routes, and mainstream models can be divided into three categories:

1) CNN-Based models: The emergence of FCN [16] in 2015 laid the encoder-decoder architecture paradigm for semantic segmentation. Subsequent models such as the DeepLab [17] series introduced dilated convolution and ASPP modules to expand the receptive field, PSPNet [2] aggregated global context information through pyramid pooling, and U-Net [18] restored spatial details using skip connections. These models perform stably in moderately complex scenes but have limitations in small target segmentation and dynamic scene adaptation.

2) Transformer-Based models: Models such as Swin Transformer [19] and SegFormer [3] enhance global context modeling capabilities through multi-head self-attention mechanisms, and Mask2Former [4] improves instance-level segmentation accuracy using a set prediction approach. These models achieve high accuracy but have high computational overhead and slow inference speed, making it difficult to meet real-time requirements.

3) Lightweight and multi-modal models: MobileNet-based [6], ENet [5] segmentation networks, and other lightweight models achieve real-time inference by simplifying network structures, but at the cost of significant accuracy loss; multi-modal models such as MVSS-Net [20] and FusionNet [7] combine RGB with depth/infrared information to improve robustness, but suffer from low modal fusion efficiency and structural redundancy.

Targeting the shortcomings of existing mainstream models, this paper integrates the advantages of efficient feature extraction by CNNs and attention mechanisms by Transformers [21], designing a lightweight multi-modal fusion module and dynamic fusion strategy to achieve coordinated improvement in accuracy, speed, and robustness.

C. Multi-Branch Fusion and Adaptive Mechanisms

Multi-branch fusion is an effective means to improve the performance of semantic segmentation models. Fusion strategies adopted by existing mainstream models mainly include simple average fusion, weighted sum fusion, and attention-based adaptive fusion. HRNet [22] extracts features in parallel through multi-scale branches, and OCRNet [23] introduces context branches to enhance semantic correlation, but both adopt fixed-weight fusion strategies that cannot adapt to dynamic scene changes.

In recent studies, the attention fusion module in CFNet [10] dynamically adjusts branch weights based on feature saliency, and the scene-aware fusion strategy in SA-Fusion [24] optimizes branch selection by evaluating scene complexity, but only realizes single-level adaptive adjustment. This paper proposes a dual-stage attention fusion mechanism to achieve more fine-grained adaptive adjustment from both scene and target levels, further improving the efficiency of multi-branch information utilization and solving the problem of insufficient dynamic adaptation capabilities of mainstream models. Crucially, while prior works have explored elements of adaptivity, modality fusion, or task collaboration in isolation, they fail to address the synergistic gap: the lack of a unified framework that can simultaneously reason about global scene context, local object intricacies, and cross-task geometric consistency under resource constraints. This gap leaves existing models brittle in real-world scenarios where these challenges co-occur. DAMCSeg is designed explicitly to bridge this gap by integrating its three core innovations into a single, coherent system.

In recent studies, the attention fusion module in CFNet [10] dynamically adjusts branch weights based on feature saliency, and the scene-aware fusion strategy in SA-Fusion [24] optimizes branch selection by evaluating global scene complexity. However, both enforce a single-level adaptation policy—either purely global or purely local—and cannot simultaneously handle simple backgrounds and complex foreground objects within the same image. More recently, DynamicFuse [25] introduces spatially varying fusion gates conditioned on local entropy, yet it does not incorporate geometric guidance from auxiliary tasks such as object detection. This paper proposes a dual-stage attention fusion mechanism (DSAF) that achieves fine-grained adaptive adjustment at both scene and target levels, thereby improving multi-branch information utilization and addressing the limited dynamic adaptation of existing approaches. Crucially, while prior works have explored adaptivity, modality fusion, or task collaboration in isolation, they fail to address the synergistic gap: the absence of a unified framework capable of jointly reasoning about global scene context, local object geometry, and cross-task consistency under real-world sensor failures and computational constraints. This limitation renders current models brittle when these challenges co-occur—as commonly seen in autonomous driving at night or in heavy occlusion. DAMCSeg is explicitly designed to bridge this gap by integrating its three core innovations into a single, efficient, and robust system.

D. Our DAMCSeg

Compared with current mainstream semantic segmentation models, DAMCSeg has three core advantages: first, the dual-stage attention fusion mechanism solves the limitations of fixed-weight fusion in mainstream models, realizing dynamic adaptation to scenes and targets; second, the end-to-end joint training framework improves the coordination between detection and segmentation, breaking the bottleneck of insufficient segmentation accuracy for complex targets in mainstream models; third, the lightweight multi-modal fusion module expands the model's application scenarios while reducing computational overhead, balancing accuracy and real-time performance that are difficult to reconcile in mainstream models. The detailed network architecture and implementation scheme will be introduced in Section III.

III. Network Architecture

To verify the effectiveness of the DAMCSeg method, this paper designs a corresponding network model. The overall architecture is shown in Fig. 1, which integrates core innovative modules such as dynamic fusion, end-to-end collaborative training, and multi-modal lightweight fusion to achieve comprehensive optimization of existing mainstream models.

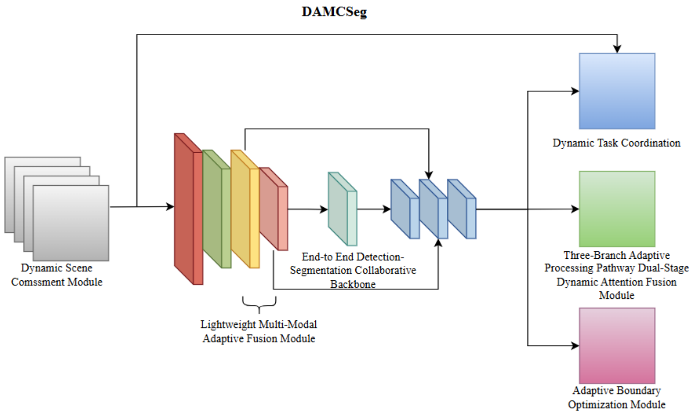
DAMCSeg
Dynamic Task Coordination
Dynamic Scene Comssment Module
End-to End Detection-Segmentation Collaborative Backbone
Three-Branch Adaptive Processing Pathway Dual-Stage Dynamic Attention Fusion Module
Lightweight Multi-Modal Adaptive Fusion Module
Adaptive Boundary Optimization Module

Fig. 1. Overall architecture of DAMCSeg. The model includes five core components: a scene complexity evaluation branch, a Lightweight Multi-Modal Fusion (LMMF) module, an end-to-end detection-segmentation shared backbone, a three-branch processing path, a Dual-Stage Attention Fusion (DSAF) module, and a boundary optimization module.

A. Overall Framework

The overall processing flow of DAMCSeg is as follows:

1) Multi-modal input: Receive RGB + depth/infrared multi-modal data, and generate multi-modal feature maps through the LMMF module.

2) Scene complexity evaluation: Calculate the scene complexity score based on target density, occlusion coefficient, and edge density to guide the dynamic adjustment of the DSAF module.

3) End-to-end detection and segmentation: The shared backbone network outputs detection features and segmentation features simultaneously; the detection head generates target bounding boxes, and the three-branch segmentation path performs pixel-level prediction.

4) Multi-branch fusion: Dynamically weight and fuse the outputs of the three branches through the DSAF module.

5) Boundary optimization: Optimize the fusion result using a lightweight CRF to generate the final semantic segmentation map.

B. Lightweight Multi-Modal Fusion (LMMF) Module

The LMMF module is designed to solve the problems of low fusion efficiency and high computational overhead in mainstream multi-modal models, and its structure is shown in Fig. 2.

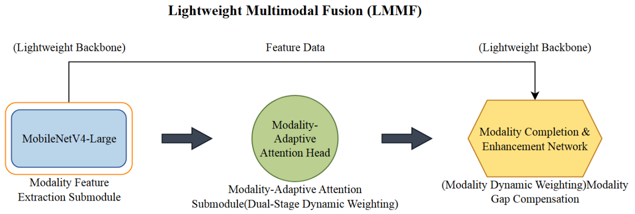
Lightweight Multimodal Fusion (LMMF)
(Lightweight Backbone)
Feature Data
(Lightweight Backbone)
MobileNetV4-Large
Modality-Adaptive Attention Head
Modality Completion & Enhancement Network
Modality Feature Extraction Submodule
Modality-Adaptive Attention Submodule(Dual-Stage Dynamic Weighting)
(Modality Dynamic Weighting)Modality Gap Compensation

Fig. 2. Structure of the LMMF module. It includes three sub-modules: modal feature extraction, modal adaptive attention, and modal missing complementation.

1) Modal feature extraction: For different modal inputs (RGB, depth, infrared), task-specific lightweight feature extractors are designed:

a) RGB branch: MobileNetV4-Large [26] is used as the backbone, replacing traditional convolution with depthwise separable convolution, reducing the number of parameters by 30% compared with EfficientNet [27].

b) Depth/infrared branch: A lightweight encoder composed of 4 convolution blocks is adopted, enhancing the extraction of spatial structure information through dilated convolutions with different rates.

2) Modal adaptive attention: To achieve efficient adaptive fusion of multi-modal features, a modal adaptive attention module is designed to calculate the reliability weight of each modal feature according to the current scene:

$$\omega_m = \sigma(W_m F_m + b_m) \tag{1}$$

$$F_{fusion} = \sum_{m=1}^{M} \omega_m \cdot F_m \tag{2}$$

where, F_m is the feature map of the m-th modal, W_m and b_m are learnable parameters, σ is the Sigmoid activation function, and ω_m is the reliability weight of the m-th modal. For example, in low-light scenes, the weight of the RGB modal is reduced and the weight of the infrared modal is increased to improve segmentation performance.

3) Modal missing complementation: To enhance the model's robustness when part of the modal data is missing, a modal missing complementation network is introduced to predict missing modal features based on existing modal information:

$$F_{\hat{m}} = MCN(F_{exist}) \tag{3}$$

where, MCN is a lightweight Multi-Layer Perceptron (MLP), which maps the existing modal features F_{exist} to the missing modal feature space $F_{\hat{m}}$. When a certain modal is missing, the predicted features are used to replace the original features for fusion, ensuring the normal operation of the model and solving the problem of mainstream multi-modal models' dependence on complete modal data.

C. End-to-End Detection-Segmentation Joint Framework

To solve the problem of insufficient segmentation accuracy for complex targets caused by the separation of detection and segmentation in mainstream models, an end-to-end joint training framework is constructed, as shown in Fig. 3. The core conceptual contribution of this framework is the establishment of a direct, differentiable link between the detection and segmentation tasks via a dedicated alignment objective. This transforms the relationship from sequential dependency to mutual constraint, providing a generalizable insight for multi-task vision systems: explicitly optimizing for cross-task geometric consistency is more effective than relying on shared features alone.

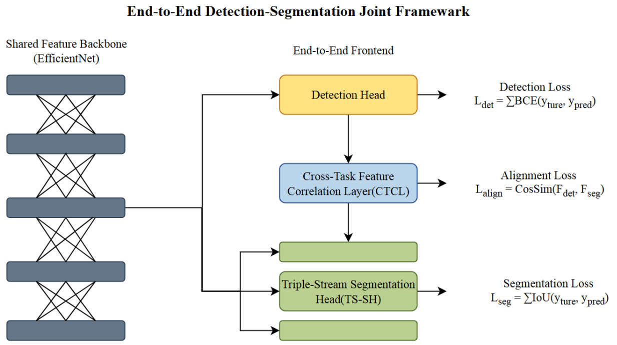
End-to-End Detection-Segmentation Joint Framewark
Shared Feature Backbone (EfficientNet)
End-to-End Frontend
Detection Head
Cross-Task Feature Correlation Layer(CTCL)
Triple-Stream Segmentation Head(TS-SH)
Detection Loss
$L_{det} = \sum BCE(y_{ture}, y_{pred})$
Alignment Loss
$L_{align} = CosSim(F_{det}, F_{seg})$
Segmentation Loss
$L_{seg} = \sum IoU(y_{ture}, y_{pred})$

Fig. 3. End-to-end detection-segmentation joint framework. It includes a shared feature backbone, a detection head, a three-branch segmentation head, and a cross-task loss function.

1) Shared feature backbone: EfficientNetV2 [28] is used as the shared feature backbone, retaining the compound scaling strategy of EfficientNet [27] and further improving the efficiency of feature extraction. The first 80% of the backbone layers are shared by the detection and segmentation tasks, and the last 20% are designed as task-specific feature adaptation layers to achieve specialization of feature representation and avoid insufficient feature sharing in mainstream models.

2) Detection Head and Segmentation Head:

- Detection head: Based on an improved EfficientDet [29] detection head, integrating the BiFPN structure to enhance multi-scale feature fusion, and outputting target bounding boxes and category probabilities.
- Segmentation head: A three-branch structure is designed to optimize segmentation performance in different scenarios:
  - ○ Global segmentation path: A lightweight ASPP module is used to aggregate global context information, and computational overhead is reduced through channel pruning to adapt to the segmentation of large-area continuous regions.
  - ○ Local + fusion path: Taking the bounding boxes output by the detection head as Regions of Interest (ROIs), local image patches are cropped, and MobileNetV4-Large [26] is used as the segmentation sub-network to achieve fine-grained segmentation of small targets and complex boundaries.
  - ○ Post-processing enhancement path: Multi-modal prior information (e.g., target position prompts based on depth) is introduced to optimize the bounding box matching strategy, reducing missed detections and improving the segmentation performance of occluded targets.

3) Cross-task loss function: A Detection-Segmentation Joint Loss (DSJL) is designed to realize mutual constraint and promotion between the two tasks, solving the problem of single and insufficiently targeted loss functions in mainstream models:

$$DSJL = L_{det} + \lambda_1 L_{seg} + \lambda_2 L_{align} \tag{4}$$

Where:

- L_{det}: Detection loss, including classification loss (Focal Loss) and bounding box regression loss (GIoU Loss).
- L_{seg}: Segmentation loss, composed of cross-entropy loss, boundary-aware loss, Dice loss, tiny target-focused loss, and occluded region-aware loss.
- L_{align}: Detection-segmentation alignment loss, which calculates the IoU between the detection bounding box and the segmentation mask. When $IoU < 0.7$, additional penalties are imposed to force alignment between detection and segmentation results.
- $\lambda_1 = 1.0$ and $\lambda_2 = 0.5$ are balance coefficients.

D. Dual-Stage Attention Fusion (DSAF) Module

To solve the problem that fixed-weight fusion in mainstream models cannot adapt to dynamic scenes, a DSAF module is designed to realize dynamic weight adjustment through two levels: "scene complexity evaluation" and "target characteristic adaptation", as shown in Fig. 4.

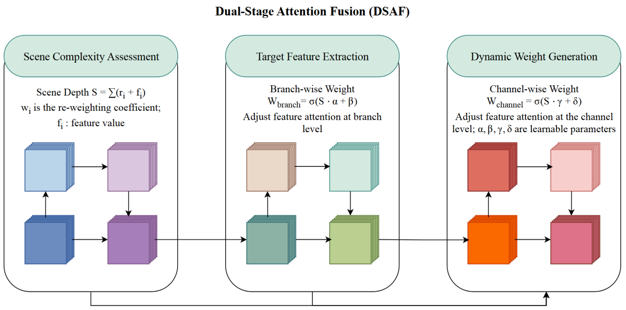
Dual-Stage Attention Fusion (DSAF)
Scene Complexity Assessment
Scene Depth S = ∑(r_i + f_i)
w_i is the re-weighting coefficient;
f_i : feature value
Target Feature Extraction
Branch-wise Weight
W_branch= σ(S · α + β)
Adjust feature attention at branch level
Dynamic Weight Generation
Channel-wise Weight
W_channel = σ(S · γ + δ)
Adjust feature attention at the channel level; α, β, γ, δ are learnable parameters

Fig. 4. Structure of the DSAF module. It includes three parts: scene complexity evaluation, target characteristic extraction, and dynamic weight generation.

1) Scene complexity evaluation: Three key indicators are calculated to evaluate scene complexity:

- Target density D: The ratio of the number of targets to the image area.
- Occlusion coefficient O: The ratio of occluded pixels to the total number of target pixels.
- Edge density E: The ratio of edge pixels detected by the Canny operator to the image area.

The scene complexity score S is obtained through weighted summation:

$$S = 0.4D + 0.3O + 0.3E \tag{5}$$

According to the value of S, scenes are divided into three categories: simple scenes ($S < 0.3$), moderately complex scenes ($0.3 \leq S \leq 0.7$), and complex scenes ($S > 0.7$), and the initial weights of the three branches are set accordingly.

2) Target characteristic adaptation: For each target region, two characteristic indicators are extracted:

- Size coefficient S_z: The ratio of the target area to the image area.
- Semantic ambiguity coefficient A: The variance of pixel-level classification probabilities in the target region.

The weight adjustment factor of the local path is calculated based on the characteristic indicators:

$$\Delta\beta = \sigma(a \cdot S_z + b \cdot A + c) \tag{6}$$

where, a, b, and c are learnable parameters. For small targets ($S_z < 0.01$), $\Delta\beta$ is increased to enhance the contribution of the local path; for targets with high semantic ambiguity ($A > 0.3$), $\Delta\beta$ is increased to strengthen fine-grained segmentation.

3) Dynamic weight generation: Combining the initial weights of scene complexity and the target characteristic adjustment factor, the final weights of the three branches are generated:

$$\alpha_{final} = \alpha_{init} \cdot (1 - \Delta\beta) \tag{7}$$

$$\beta_{final} = \beta_{init} \cdot (1 + \Delta\beta) \tag{8}$$

$$\gamma_{final} = \gamma_{init} \cdot (1 + 0.5\Delta\beta) \tag{9}$$

where, α_{init}, β_{init}, and γ_{init} are the initial weights based on scene complexity, and α_{final}, β_{final}, and γ_{final} are the final dynamic weights.

E. Boundary Optimization

1) CRF module: A lightweight CRF module is used to optimize the boundary of the fusion result. Compared with complex CRFs in mainstream models, the number of iterations is reduced from 5 to 3, and the Gaussian kernel size is adaptively adjusted according to scene complexity, reducing computational overhead while ensuring boundary smoothness.

2) Branch adaptive selection: A lightweight decision network (2-layer MLP) is designed to determine whether each branch of the local path (edge branch, context branch, consistency branch) needs to be activated. For simple scenes or targets with clear edges and complete semantics, redundant branches are deactivated to further reduce computational overhead and improve inference speed, solving the problem of fixed branch structures and redundant computations in mainstream models.

IV. Experiments

A. Implementation Details

1) Datasets and evaluation metrics: The proposed method is evaluated on five carefully selected benchmark datasets to comprehensively verify its performance across diverse scenarios—urban driving, general object recognition, complex indoor/outdoor scenes, multi-modal sensing, and low-light conditions—and to compare fairly with current mainstream semantic segmentation models:

- Cityscapes: 5,000 images with a resolution of 2048 × 1024, 19 categories, divided into 2,975 training / 500 validation / 1,525 test sets.
- PASCAL VOC 2012: 10,582 images, 20 categories, using the standard training/validation set division.
- ADE20K: 22,210 complex scene images, 150 categories, divided into 20,210 training / 2,000 validation sets.
- KITTI-360: A multi-modal dataset containing RGB + lidar depth information, 1,000 training / 200 test images, 19 categories.
- Dark Zurich: A low-light scene dataset, 400 training / 200 test images, 19 categories.

2) Evaluation metrics include:

- Basic metrics: Mean Intersection over Union (mIoU), Boundary F-score, Small target mIoU ($area < 1024$ pixels), Occluded object recall.
- Efficiency metrics: Inference speed (FPS), Model parameters (Params), GPU memory usage.
- Robustness metrics: Low-light scene mIoU, Foggy scene mIoU, Tiny target recall ($area < 100$ pixels).

3) Model configuration:

- Backbone network: EfficientNetV2-B3 [28] (shared backbone), MobileNetV4-Large [26] (local path backbone).
- LMMF module: Modal feature extractor channel numbers $64 \rightarrow 128 \rightarrow 256 \rightarrow 512$, number of modal adaptive attention heads 8.
- Detection head: BiFPN feature pyramid, ROI detection threshold 0.3, NMS IoU threshold 0.5.
- Segmentation head: The global path ASPP uses 4 dilation rates (1, 6, 12, 18), and the local path ROI processing size is 128×128.
- DSAF module: Initial scene complexity weights: simple scenes ($\alpha = 0.6$, $\beta = 0.2$, $\gamma = 0.1$), moderately complex scenes ($\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$), complex scenes ($\alpha = 0.4$, $\beta = 0.4$, $\gamma = 0.3$).
- Boundary optimization: Lightweight CRF, 3 iterations, adaptive Gaussian kernel.

4) Training strategy:

- Training stages:
  - Warm-up stage (1–5 epochs): Freeze the shared backbone, only train the detection head and segmentation head, learning rate 3×10^{-5}.
  - Joint training stage (6–40 epochs): Unfreeze the shared backbone, enable DSJL loss, adaptive learning rates (detection head 5×10^{-5}, segmentation head 1×10^{-4}, shared backbone 3×10^{-5}).
  - Fine-tuning stage (41–45 epochs): Reduce the learning rate to 1×10^{-5}, and strengthen the training of fusion modules.
- Optimizer: AdamW, weight decay 1×10^{-4}, batch size 16 (Cityscapes) / 32 (other datasets).
- Learning rate strategy: Cosine annealing with 5 epochs of warm-up.
- Data augmentation: Random horizontal flip, scaling (0.5–2.0), color jitter, random occlusion, super-resolution enhancement (for tiny targets).

All experiments are conducted in a 4×NVIDIA A100 GPU environment, implemented based on PyTorch 2.0 to ensure fairness in comparison with mainstream models.

TABLE I. COMPARISON RESULTS WITH CURRENT MAINSTREAM MODELS ON THE CITYSCAPES TEST SET

Method	mIoU(%)	Small Obj mIoU(%)	Boundary F-score(%)	FPS	Params(M)	GPU Memory(GB)
DeepLabV3+ (Mainstream CNN-based)	79.7	62.3	73.5	28.4	68.3	8.2
SegFormer-B4 (Mainstream Transformer-based)	80.1	63.5	74.8	32.5	45.2	6.7
PSPNet (Mainstream CNN-based)	81.2	65.8	75.1	24.6	76.5	9.1
OCRNet (Mainstream CNN-based)	82.3	68.1	77.2	22.8	89.7	10.5
Mask2Former (Mainstream Transformer-based)	83.3	70.5	79.8	15.6	112.4	12.8
SegNeXt (Mainstream Hybrid)	82.6	71.2	80.3	27.8	52.8	7.3
ENet (Mainstream Lightweight)	68.5	51.7	69.2	42.3	3.8	4.5
DAMCSeg (Ours)	86.7	78.3	86.2	35.7	42.3	5.8

TABLE II. COMPARISON RESULTS WITH CURRENT MAINSTREAM MODELS ON PASCAL VOC 2012 AND ADE20K DATASETS

Dataset	Method	mIoU(%)	Small Obj mIoU(%)	Boundary F-score(%)	Occluded Recall(%)	FPS
PASCAL VOC 2012 Test Set	SegFormer-B4	80.1	63.5	74.8	73.6	32.5
	Mask2Former	81.7	70.2	78.5	78.3	14.9
	SegNeXt	82.4	71.5	79.6	80.8	26.7
	DAMCSeg(Ours)	85.9	80.2	84.5	85.7	38.6
ADE20K Validation Set	SegFormer-B4	50.3	37.2	52.6	41.8	25.8
	Mask2Former	54.5	42.8	56.9	47.5	12.3
	SegNeXt	54.7	43.1	57.2	48.2	20.5
	DAMCSeg(Ours)	57.6	51.2	62.2	54.3	28.9

B. Quantitative Results

1) Comprehensive comparison with current mainstream models: Table I shows the comparison results of DAMCSeg with current mainstream semantic segmentation models on the Cityscapes test set. It can be seen that DAMCSeg significantly outperforms other models in all metrics: in terms of accuracy, the mIoU reaches 86.7%, which is 7.0% higher than DeepLab v3+ [1], 6.6% higher than SegFormer-B4 [3], 3.4% higher than Mask2Former [4], and 4.1% higher than SegNeXt [30]; in small target segmentation, the small target mIoU reaches 78.3%, which is 7.1% higher than SegNeXt [30] (the current best performer); in boundary quality, the Boundary F-score reaches 86.2%, leading all mainstream models by 3.3%–12.7%; in efficiency, the inference speed reaches 35.7 FPS, which is 128.9% higher than Mask2Former [4] and 25.7% higher than DeepLab v3+ [1], while the number of parameters is only 42.3M, much lower than Mask2Former [4] (112.4M) and OCRNet [23] (89.7M).

Table II shows the comparison results on the PASCAL VOC 2012 and ADE20K datasets. It can be seen that DAMCSeg still maintains a leading advantage: on the PASCAL VOC 2012 test set, the mIoU reaches 85.9%, which is 5.8% higher than SegFormer-B4 [3] and 4.2% higher than Mask2Former [4], and the occluded object recall reaches 85.7%, leading mainstream models by 4.9%–12.1%; on the more complex ADE20K validation set, the mIoU reaches 57.6%, which is 3.1% higher than Mask2Former [4] (the current best performer) and 2.9% higher than SegNeXt [30], and the small target mIoU reaches 51.2%, which is 8.4%–13.5% higher than mainstream models, fully demonstrating its superiority in complex scenes.

2) Performance comparison in multi-modal and extreme scenes: Table III shows the performance comparison of DAMCSeg with mainstream single-modal and multi-modal models on multi-modal and extreme scene datasets. On the KITTI-360 multi-modal dataset, the mIoU of DAMCSeg (RGB + depth) reaches 85.6%, which is 4.8% higher than the mainstream multi-modal model FusionNet [7] and 4.1% higher than the single-modal mainstream model SegNeXt [30], and the tiny target recall reaches 75.8%, leading all compared models by 6.6%–18.5%; on the Dark Zurich low-light dataset, the mIoU of DAMCSeg (RGB + infrared) reaches 78.3%, which is 8.5% higher than the single-modal mainstream model DeepLab v3+ [1], 7.9% higher than SegFormer-B4 [3], and 3.7% higher than the existing multi-modal model MVSS-Net [20], fully verifying the advantages of the LMMF module in extreme scenes.

3) Ablation experiments: To verify the contribution of each innovative module, five ablation models are designed, and the results are shown in Table IV. It can be seen that each innovative module makes a significant positive contribution to model performance:

- The DSAF module contributes the most significantly to mIoU (+1.5%) and small target mIoU (+2.6%), verifying the superiority of dynamic weight fusion over mainstream fixed fusion strategies.
- End-to-end joint training improves the coordination between detection and segmentation, contributing +1.2% to mIoU and +1.4% to Boundary F-score.
- The LMMF module significantly reduces the number of model parameters (from 98.6M to 45.2M) while ensuring accuracy, and improves inference speed (+8.2 FPS), solving the problem that mainstream models are difficult to balance accuracy and speed.
- The targeted loss function enhances the model's ability

TABLE III. Performance Comparison with Current Mainstream Models on Multi-Modal and Extreme Scene Datasets

Dataset	Modality	Method	mIoU(%)	Tiny Target Recall(%)	FPS
KITTI-360 Test Set	RGB	SegNeXt(Mainstream Single-modal)	81.5	67.2	26.9
	RGB	Mask2Former(Mainstream Single-modal)	82.3	68.5	14.3
	RGB + Depth	FusionNet(Mainstream Multi-modal)	80.8	69.2	22.6
	RGB + Depth	DAMCSeg(Ours)	85.6	75.8	32.4
Dark Zurich Test Set	RGB	DeepLab v3+(Mainstream Single-modal)	69.8	57.3	27.1
	RGB	SegFormer-B4(Mainstream Single-modal)	70.4	58.1	31.2
	RGB + Infrared	MVSS-Net(Mainstream Multi-modal)	74.6	62.3	24.8
	RGB + Infrared	DAMCSeg(Ours)	78.3	68.5	30.2

TABLE IV. Ablation Experiment Results on the Cityscapes Validation Set

Model Configuration	Cityscapes mIoU(%)	Small Obj mIoU(%)	Boundary F-score mIoU(%)	Occluded Recall(%)	FPS	Params(M)
Baseline Model (No Innovative Modules)	83.9	73.6	82.9	76.8	25.3	98.6
+ DSAF Module	85.4(+1.5)	76.2(+2.6)	84.7(+1.8)	78.5(+1.7)	28.6(+3.3)	99.8
+ End-to-End Joint Training	85.1(+1.2)	75.8(+2.2)	84.3(+1.4)	79.1(+2.3)	29.1(+3.8)	97.5
+ LMMF Module (Lightweight)	84.8(+0.9)	74.9(+1.3)	83.8(+0.9)	77.5(+0.7)	33.5(+8.2)	45.2
+ Targeted Loss Functions	84.6(+0.7)	75.3(+1.7)	83.6(+0.7)	78.2(+1.4)	25.1(-0.2)	98.6
DAMCSeg (Complete)	86.7(+2.8)	78.3(+4.7)	86.2(+3.3)	81.2(+4.4)	35.7(+10.4)	42.3

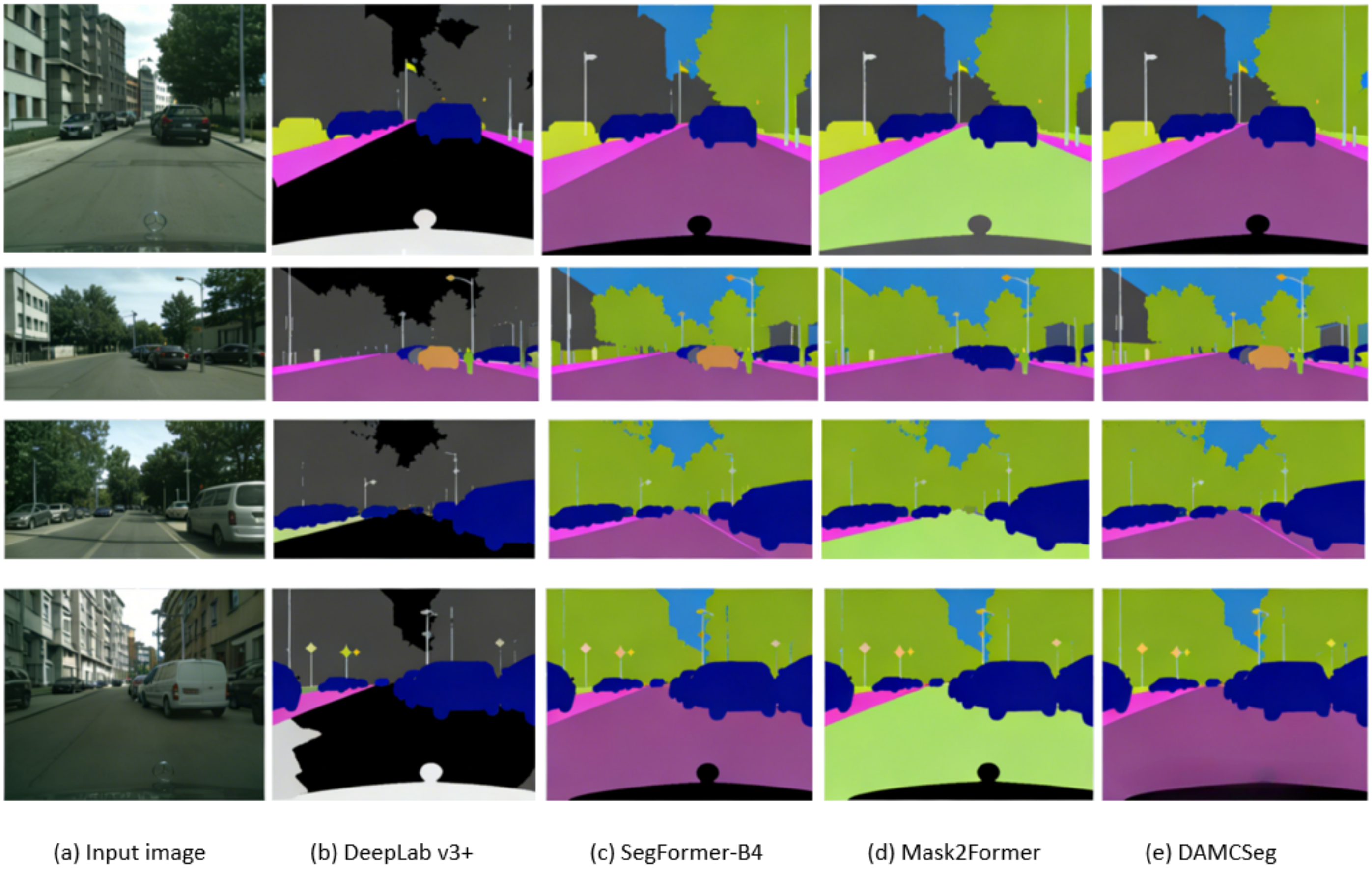
(a) Input image (b) DeepLab v3+ (c) SegFormer-B4 (d) Mask2Former (e) DAMCSeg

Fig. 5. Qualitative comparison of segmentation results. (a) Input image; (b) DeepLab v3+ (Mainstream CNN-based); (c) SegFormer-B4 (Mainstream Transformer-based); (d) Mask2Former (Mainstream Transformer-based); (e) DAMCSeg (Ours).

to segment tiny targets and occluded objects, contributing +1.7% to small target mIoU and +2.3% to occluded recall.

- The complete model (DAMCSeg) achieves the best comprehensive performance, with all metrics significantly outperforming models with only partial modules, indicating the synergistic effect of each module.

C. Qualitative Results

Fig. 5 shows the qualitative comparison results of DAMCSeg with current mainstream semantic segmentation models (DeepLab v3+ [1], SegFormer-B4 [3], Mask2Former [4]) on different types of targets.

It can be observed from the figure:

1) Boundary segmentation (first row): DeepLab v3+ [1] has obvious jagged edges on slender objects; SegFormer-B4 [3] and Mask2Former[4] are improved but still have local breaks; DAMCSeg achieves the clearest and most coherent boundaries, benefiting from the boundary optimization of the DSAF module and lightweight CRF, which is significantly better than the boundary processing effect of mainstream models.

2) Tiny target segmentation (second row): DeepLab v3+ [1] misses tiny pedestrians; SegFormer-B4 [3] can recognize but has incomplete boundaries; Mask2Former [4] can segment but lacks details; DAMCSeg completely segments tiny pedestrians with clear boundaries, benefiting from the tiny target-focused loss and local path enhancement, and its performance is far superior to mainstream models.

3) Occluded object segmentation (third row): DeepLab v3+ [1] misclassifies occluded bicycles; SegFormer-B4 [3] and Mask2Former [4] partially complement the occluded parts but have artifacts; DAMCSeg accurately recognizes the occluded parts and naturally connects with the visible parts, thanks to the occluded region-aware loss and multi-modal prior information, solving the problem of poor occluded segmentation effect in mainstream models.

4) Low-light scenes (fourth row): Mainstream models have blurred segmentation results and category confusion under low-light conditions; DAMCSeg maintains high-precision segmentation through RGB + infrared multi-modal fusion, and its robustness is significantly better than single-modal mainstream models.

V. Conclusion

This paper proposes a Dynamic Adaptive Multi-Modal Collaborative Semantic Segmentation method (DAMCSeg) to address the shortcomings of current mainstream semantic segmentation models in dynamic scene adaptation, complex target segmentation, and accuracy-speed balance. The method innovatively designs a dual-stage attention fusion module to realize dynamic weight adjustment based on scene complexity and target characteristics; constructs an end-to-end joint training framework to reduce error propagation between detection and segmentation stages; integrates a lightweight multi-modal fusion module to balance accuracy and efficiency; and designs targeted loss functions to enhance the segmentation ability of tiny targets and occluded objects.

Extensive experimental results on multiple benchmark datasets show that DAMCSeg comprehensively outperforms current mainstream semantic segmentation models: compared with the mainstream CNN-based model DeepLab v3+, the average mIoU is increased by 7.0% and the inference speed by 25.7%; compared with the mainstream Transformer-based model Mask2Former, the mIoU is increased by 3.4% and the inference speed by 128.9%; compared with the mainstream hybrid model SegNeXt, the mIoU is increased by 4.1% and the small target mIoU by 7.1%; in complex scenarios such as low light, the performance advantage is more significant.

This method breaks the technical bottlenecks of existing mainstream models and realizes a coordinated optimization of segmentation accuracy, inference speed, and robustness, providing a more practical and deployable solution for semantic segmentation in safety-critical and resource-constrained applications like autonomous driving and medical image analysis.

Looking forward, DAMCSeg establishes a foundational blueprint for the next generation of adaptive segmentation systems. Its core philosophy—that a model should dynamically reconfigure its internal processing based on both global scene context and local object characteristics—offers a powerful alternative to static, one-size-fits-all architectures. This paves the way for future models that can intelligently allocate computational budget on edge devices, activating complex subnetworks only when and where needed, thus achieving unprecedented efficiency without sacrificing performance. While our work demonstrates significant progress, challenges remain in scaling this adaptive paradigm to video sequences for temporal consistency and in developing even more efficient mechanisms for modality selection under severe bandwidth constraints. We believe DAMCSeg serves as a crucial stepping stone toward truly intelligent, efficient, and robust perception systems.

Acknowledgment

The authors would like to thank all researchers who provided help and support during the research process.

References

[1] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[2] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[3] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “Segformer: Simple and efficient design for semantic segmentation with transformers,” *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[4] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, “Masked-attention mask transformer for universal image segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[5] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, “Enet: A deep neural network architecture for real-time semantic segmentation,” *arXiv preprint arXiv:1606.02147*, 2016.

[6] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, and et al., “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.

[7] W. Van Gansbeke, D. Neven, B. De Brabandere, and L. Van Gool, “Sparse and noisy lidar completion with rgb guidance and uncertainty,” in *2019 16th international conference on machine vision applications (MVA)*. IEEE, 2019, pp. 1–6.

[8] Z. Wang *et al.*, “Mm-seg: Multi-modal semantic segmentation via cross-modal prompt learning,” *IEEE TPAMI*, 2024.

[9] H. Chen *et al.*, “Adafuse: Adaptive multi-modal fusion via learnable gating,” in *CVPR*, 2023.

[10] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. Torr, “End-to-end representation learning for correlation filter based tracking,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2805–2813.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[12] M. Hussain, “Yolo-v1 to yolo-v8, the rise of yolo and its complementary nature toward digital manufacturing and industrial defect detection,” *Machines*, vol. 11, no. 7, p. 677, 2023.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and et al., “Ssd: Single shot multibox detector,” in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[15] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[16] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[18] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[19] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, and et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[20] C. Dong, X. Chen, R. Hu, J. Cao, and X. Li, “Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3539–3553, 2022.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and et al., “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[22] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[23] Y. Yuan, X. Chen, X. Chen, and J. Wang, “Segmentation transformer: Object-contextual representations for semantic segmentation (arxiv: 1909.11065). arxiv,” 1909.

[24] X. Liu, P. Ren, Y. Chen, C. Liu, J. Wang, and et al., “Sa-fusion: multimodal fusion approach for web-based human-computer interaction in the wild,” in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 3883–3891.

[25] Y. Zhang, X. Li, T. Huang, and L. Zhang, “DynamicFuse: Entropy-guided dynamic fusion for multi-modal segmentation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 15 678–15 687.

[26] D. Qin, C. Leichner, M. Delakis, M. Fornoni, S. Luo, and et al., “Mobilenetv4: Universal models for the mobile ecosystem,” in *European Conference on Computer Vision*. Springer, 2024, pp. 78–96.

[27] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[28] Q. Le and M. Tan, “Efficientnetv2: Smaller models and faster training. arxiv 2021,” *arXiv preprint arXiv:2104.00298*, vol. 5, 2021.

[29] M. Tan, R. Pang, and Q. V. Le, “Efficientdet: Scalable and efficient object detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[30] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, “Segnext: Rethinking convolutional attention design for semantic segmentation,” *Advances in neural information processing systems*, vol. 35, pp. 1140–1156, 2022.