Scientific Reports

https://doi.org/10.1038/s41598-026-39249-1

Article in Press

Dynamic background motion object semantic segmentation algorithm based on generative adversarial network and transformer collaboration

Received: 10 September 2025

Accepted: 3 February 2026

Published online: 08 March 2026

Cite this article as: Li Y., Luo Z., Chen T. *et al.* Dynamic background motion object semantic segmentation algorithm based on generative adversarial network and transformer collaboration. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-39249-1

YiQiang Li, ZhenBao Luo, Tao Chen, XinJun Huang, ChaoZe Zhong, Ge Zhu, DaiZhong Jin, Chen Cheng, Yi Zhang, JingTong Zhao & PengCheng Gao

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

ARTICLE IN PRESS

Dynamic Background Motion Object Semantic Segmentation Algorithm Based on Generative Adversarial Network and Transformer Collaboration

YiQiang Li[1,*], ZhenBao Luo[1], Tao Chen[1], XinJun Huang[2], ChaoZe Zhong[1], Ge Zhu[1], DaiZhong Jin[1], Chen Cheng[1], Yi Zhang[1], JingTong Zhao[1], PengCheng Gao[1],
1. Norla Institute of Technical Physics, No. 7, Section 4, Renmin South Road, Wuhou District, Chengdu City, Sichuan Province, 610095, China.
2, Jiangxi Hongdu Aviation Industry Group Co., Ltd. Building 16, Zone 8, Hongdu Aviation Industry Group, Xinxiqiao Road, Qingyunpu District, Nanchang City, Jiangxi Province, 330024, China.
* Corresponding author: YiQiang Li, E-mail: 18883383440@163.com

Abstract: Semantic segmentation of moving objects in dynamic backgrounds faces core challenges such as background interference and blurred target features. This study proposes an innovative architecture that integrates Generative Adversarial Network (GAN) with Transformers. The GAN module enhances adaptability to dynamic backgrounds through adversarial training, while the self-attention mechanism in the Transformer captures long-range semantic dependencies. A gated fusion strategy is designed to achieve dynamic balancing of multimodal features. The method employs a conditional GAN to generate dynamic background samples with variations in illumination and motion blur. A Transformer-based encoder-decoder structure is used to model global contextual relationships. A temporal attention module is introduced to incorporate motion vector fields, improving temporal consistency. Additionally, a KL-divergence (KL) constrained semantic consistency loss optimizes the plausibility of generated samples. Experiments are conducted on both a multi-dimensional simulated dataset and the real-world KITTI dataset. Results show that the proposed model achieves an average Intersection over Union (IoU) of 85.6% in standard dynamic scenes, outperforming DeepLabv3+ by 9.2 percentage points. In low-light and high-speed motion scenarios, the robustness index reaches 92.0%, 8.5 points higher than baseline models. Ablation studies demonstrate that removing the Transformer leads to a 6.7% drop in mIoU, while excluding the feature fusion module reduces robustness by 4.0%, confirming the necessity of both components. Temporal analysis reveals that the model maintains a stable performance of 84.5%–86.5%

ARTICLE IN PRESS

over 20-frame sequences, with fluctuation reduced by 63% compared to baseline. The adversarial training improves the model's adaptability to lighting changes by 5.3%. The multi-head self-attention (MSA) mechanism reduces long-range misclassification by 6.7%. The gated fusion strategy lowers false positive rates in background-disturbed regions by 12.8%. This framework optimizes segmentation through a generator-segmenter feedback loop, effectively balancing dynamic background noise suppression and semantic fidelity. The contributions are threefold: (1) The first semantic segmentation framework to deeply integrate GANs and Transformers. (2) A theoretical model for dynamic feature gating and semantic consistency constraints. (3) A standardized evaluation system covering 10 dynamic background types and five illumination gradients. This study provides key technical support for real-time environmental perception in autonomous driving and intelligent surveillance, advancing both the theoretical and practical frontiers of dynamic scene understanding.

Keywords: Dynamic background; Semantic segmentation; Transformer; Generative adversarial network; Dynamic characteristics

1. Introduction

Semantic segmentation of moving objects in dynamic backgrounds aims to accurately identify and segment different categories of moving targets from complex and changing scenes. It is a key research area in the field of computer vision [1]. This technique has broad practical applications in domains such as intelligent security, traffic surveillance, and video analysis. For example, semantic segmentation of moving objects in surveillance footage can enable the detection and early warning of abnormal behaviors. In autonomous driving, precise segmentation of moving targets helps vehicles perceive their surroundings in real time and make informed decisions [2].

However, the complexity of dynamic backgrounds presents significant challenges for semantic segmentation of moving objects. Factors such as illumination changes, object motion, and background disturbances can cause interference between background and foreground information, making it difficult for models to accurately distinguish targets from the background [3]. Traditional semantic segmentation methods primarily rely on handcrafted features, such as color and texture, which exhibit poor robustness under dynamic conditions and struggle to adapt to complex scene variations [4]. With the advancement of deep learning, convolutional neural network (CNN)-based methods have achieved substantial progress in semantic segmentation by automatically

ARTICLE IN PRESS

learning deep image features, demonstrating strong performance in static background scenarios. Nevertheless, existing methods still face limitations in dynamic backgrounds [5]. On one hand, the local receptive field of CNN restricts their ability to capture long-range contextual dependencies, making it difficult to model inter-object relationships in dynamic scenes. On the other hand, current methods lack effective modeling of background dynamics, leading to degraded segmentation accuracy when background interference is strong [6]. Generative Adversarial Network (GAN), through adversarial training between a generator and a discriminator, can produce realistic image samples. It shows strong capabilities in data augmentation and image generation tasks. Applying GAN to dynamic background segmentation enables the generation of diverse dynamic samples, enriching the training set and improving model adaptability. Meanwhile, Transformers, leveraging self-attention mechanisms, can effectively capture global contextual dependencies and have achieved impressive results in both natural language processing and computer vision. Introducing Transformers into semantic segmentation tasks can overcome the limitations of CNNs' local receptive fields and enhance the modeling of long-range contextual information [7].

Based on the above analysis, this study innovatively proposes a dynamic background moving object semantic segmentation algorithm deeply integrating GAN and Transformers, aiming to break through the core limitations of existing methods. Instead of simply superimposing the functions of the two models, this algorithm enables the dynamic background modeling capability of GAN and the global semantic capture advantage of Transformers to form a complementary and synergistic effect through architectural-level collaborative design. The GAN module generates background samples containing complex dynamic features such as illumination variations and motion blur via conditional regularization. It enriches the diversity of training data and endows the model with adaptive capacity against dynamic interference. In contrast, the Transformer breaks through the local receptive field limitation of CNN by virtue of the multi-head self-attention mechanism, accurately captures the long-range semantic correlations of moving objects, and solves the problem of semantic confusion between objects and backgrounds in dynamic scenarios. Particularly crucially, the designed gated fusion strategy breaks the simple serial mode of existing Transformer-GAN hybrid models. It can adaptively adjust the weight ratio of background features and semantic features according to the dynamic characteristics of scenarios. It strengthens the anti-interference capability of GANs in regions with severe background

ARTICLE IN PRESS

interference, and highlights the semantic accuracy of Transformers in object boundary regions, thus achieving a dynamic balance between dynamic features and semantic features. Meanwhile, the embedded temporal attention module and the semantic consistency loss constrained by KL divergence further make up for the deficiencies of existing research in spatiotemporal continuity modeling and semantic rationality of generated samples, effectively improving the segmentation stability and robustness of the model in complex dynamic scenarios. This method provides a brand-new technical path and theoretical support for semantic segmentation in dynamic environments, and has both theoretical innovation value and application potential in practical scenarios such as intelligent surveillance and autonomous driving.

2. Literature Review

In the field of moving object semantic segmentation under dynamic backgrounds, researchers have continuously explored segmentation accuracy and robustness in complex scenes, driven by advancements in computer vision. He et al. (2021) addressed model diagnosis and optimization for movable object segmentation in autonomous driving and proposed a visual analysis method, Visual Analytics for Semantic Segmentation (VASS). This method learned context-aware spatial representations to extract key spatial features and generates adversarial samples to evaluate spatial robustness. It provided an interpretable framework for enhancing segmentation of critical targets such as cargo and pedestrians and effectively identifies weak segmentation regions in typical traffic scenes [8]. For moving object segmentation using 3D LiDAR data, Chen et al. (2021) were the first to employ sequential range images from rotating LiDAR sensors. They built a CNN-based learning framework that modeled temporal data to distinguish between moving and static objects in dynamic scenes. Using the SemanticKITTI dataset, they established the first LiDAR-based benchmark for moving object segmentation, offering a standardized platform for future studies [9]. Kim et al. (2022) further improved LiDAR-based segmentation by introducing RVMOS, a Range-View framework that fused semantic and motion features. By designing a feature extraction module suited for range views and incorporating temporal modulation strategies, the method significantly enhanced segmentation accuracy while maintaining computational efficiency. It achieved a 19% improvement in mean IoU over state-of-the-art methods on the SemanticKITTI benchmark, enabling real-time performance in dynamic environments [10]. In unsupervised learning, Bielski and Favaro (2022) proposed the MOVE method, which utilized local motion of foreground objects to generate realistic images. This

ARTICLE IN PRESS

allowed training segmentation models on unlabeled data, addressing the reliance of traditional supervised learning on large annotated datasets. MOVE achieved state-of-the-art performance in unsupervised salient object detection and segmentation, offering a novel self-supervised approach for dynamic scenes [11]. In the same period, Mersch et al. (2022) introduced a sparse 4D convolutional network to segment retreating moving objects in LiDAR data. By integrating sparse features across time, their model improved trajectory capture capability, providing technical support for analyzing point cloud sequences in complex dynamic environments [12]. From an application-oriented perspective, Dang and Bui (2023) developed a multi-scale fully convolutional network to enhance semantic segmentation accuracy for mobile robot navigation. By extracting multi-resolution features, the method improved recognition of traversable areas and obstacles in real-world navigation scenarios, providing reliable semantic cues for path planning [13]. Manakitsa et al. (2024) conducted a comprehensive review of machine learning and deep learning applications in machine vision, outlining developments in object detection, semantic segmentation, and action recognition. They highlighted multimodal fusion, lightweight model design, and unsupervised learning as key future research directions, offering valuable cross-domain theoretical insights [14]. To meet the high-precision perception demands of autonomous driving, Song et al. (2024) proposed SSF-MOS, a semantic scene flow-assisted framework for moving object segmentation. By integrating semantic features of dynamic point clouds with motion vectors from scene flow, this method addressed the disconnect between semantic and motion features in conventional approaches and significantly improved spatiotemporal consistency in segmenting dynamic objects such as vehicles and pedestrians in complex traffic environments [15]. In the latest study, Tang et al. (2025) tackled boundary ambiguity in 3D LiDAR point cloud segmentation. They introduced gradient enhancement techniques and motion consistency constraints to improve boundary sharpness and ensure spatiotemporal continuity of motion features across adjacent frames. Their method further enhanced the integrity and accuracy of moving object contours, offering new strategies for fine-grained point cloud processing [16].

In the field of semantic segmentation of moving objects in dynamic backgrounds and related applications, research in recent years has continuously focused on scenario adaptability and task-specific optimization. Cheng and Zheng (2022) [17] proposed a sequential semantic segmentation method for road contours to meet the requirements of intelligent transportation scenarios, providing

ARTICLE IN PRESS

accurate semantic support for path planning and speed decision-making. It highlighted the practical value of sequential segmentation in dynamic traffic environments. Acharya et al. (2022) [18] approached the problem from the perspective of domain adaptation, combining 3D models, hierarchical edge maps, and semantic segmentation to address the cross-domain adaptation issue in single-image localization, thus expanding the application boundaries of semantic segmentation in multi-modal fusion scenarios. To tackle the label efficiency challenge in video object segmentation, Lu et al. (2023) [19] introduced motion cue-enhanced feature representation, which reduced label dependency while improving the accuracy of dynamic object segmentation, offering an effective approach for feature utilization in the temporal dimension. Gupta and Kumar (2025) [20] proposed a combined framework of SemSegX and TripForceNet for surveillance scenarios, achieving accurate tracking and segmentation of moving objects and enhancing the practical performance of semantic segmentation technology in the security and surveillance field. These studies have advanced the development of the relevant field from the perspectives of application scenario adaptation, cross-domain optimization, and temporal feature utilization, respectively, and also provided multi-dimensional technical references for semantic segmentation in dynamic backgrounds.

Based on existing research progress, three key research gaps remain in moving object segmentation under dynamic backgrounds: 1. Transformer-based segmentation techniques lack data augmentation support for dynamic backgrounds, resulting in limited generalization ability. 2. Transformer-GAN hybrid models fail to achieve deep collaboration, making it difficult to balance background modeling and target semantic extraction. 3. There is a lack of spatiotemporal feature fusion and semantic consistency constraint mechanisms for dynamic scenes. To address the above gaps, this study proposes a semantic segmentation architecture with deep collaboration between GAN and Transformer. Its innovative positioning is reflected in three aspects: 1. Breaking the simple serial mode of existing hybrid models, a gated fusion strategy is designed to achieve adaptive balance between dynamic background features and target semantic features. 2. A semantic consistency loss constrained by KL divergence is introduced to solve the problem of semantic rationality of GAN-generated samples. 3. A temporal attention module is embedded to integrate motion vector fields, improving the temporal consistency of dynamic targets. Through the above designs, this study realizes the collaborative optimization of dynamic background modeling, global semantic extraction, and

ARTICLE IN PRESS

spatiotemporal feature fusion. It provides a new technical path for moving object segmentation in complex dynamic scenes.

3. The Model Design of Cooperative Transformer for GAN

3.1 Basic Model Design

(1) Design of GAN model

The GAN module is designed to enhance the model's adaptability to complex dynamic scenes by generating diverse dynamic background samples through adversarial training. This module consists of a generator G and a discriminator D, which are optimized through adversarial learning. The generator adopts a multi-layer cascaded convolutional neural network (CNN) architecture. Its input is a fused vector of a random noise vector z and static background features B_{static}. Through a series of deconvolution operations, it progressively upsamples the input to generate high-resolution dynamic background images $G(z,B_{static})$. Conditional regularization layers (e.g., conditional batch normalization) are incorporated into the network to ensure that the generated dynamic backgrounds contain designated characteristics such as lighting variation, motion blur, and background object perturbation. The discriminator employs a deep convolutional architecture, taking either real dynamic background samples B_{real} or generated samples $G(z, B_{static})$ as input. It extracts spatial features through multiple convolutional layers and outputs a binary probability to determine the authenticity of the input.

During adversarial training, the generator aims to minimize the distributional difference between generated and real samples, while the discriminator seeks to maximize the probability of correctly classifying real samples. The loss function adopts an improved Wasserstein distance with a gradient penalty term to ensure training stability. The specific form is as follows:

$$L_{GAN} = E_{B_{real}\sim p_{data}}[D(B_{real})] - E_{z,B_{static}\sim p_z\times p_{static}}[D(G(z,B_{static}))] + \lambda E_{\hat{B}\sim p_{\hat{B}}}[(\| \nabla_{\hat{B}} D(\hat{B}) \|_2 -1)^2] \quad (1)$$

p_{data} is the real dynamic background distribution. p_z and p_{static} are the prior distributions of noise and static background. $\hat{B}$ is the linear interpolation between real samples and generated samples. λ is the gradient penalty coefficient [21]. Through the above design,

the generator can output the dynamic background close to the real scene and provide rich training data for the subsequent semantic segmentation model [22]. In Figures 1 and 2, the structural design of the GAN model is shown.

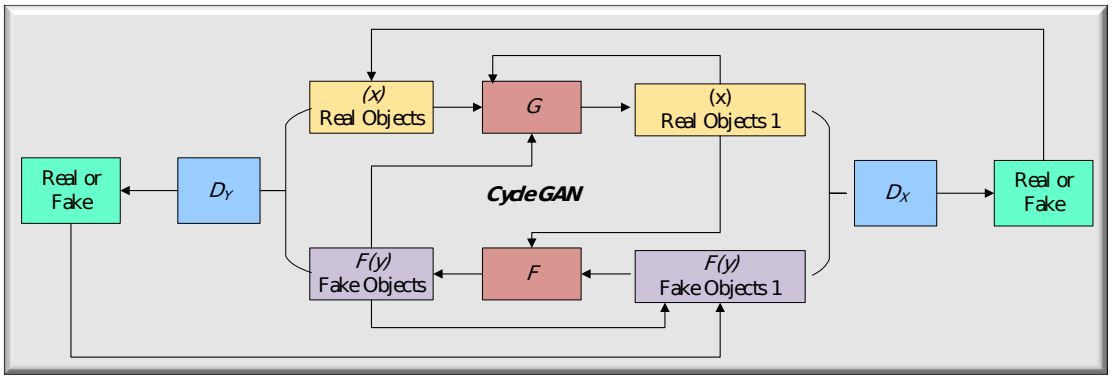


Figure 1 structural design of GAN model

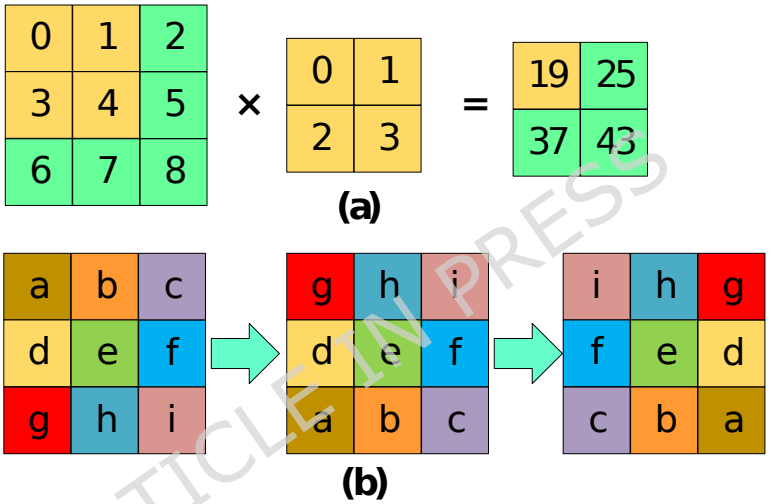


Figure 2 Crossover algorithm

(2) Transformer model design

The core of the Transformer module lies in leveraging the self-attention mechanism to capture long-range contextual dependencies, enabling accurate extraction of semantic features of moving objects [23]. This module adopts an encoder–decoder architecture. The input is an image feature map $F \in R^{H\times W\times C}$, which is preprocessed by convolution. The feature map is first divided into $N = H \times W/P^2$ non-overlapping patches of size $P \times P$, and each patch is flattened into a vector of dimension $C' = P^2 \times C$, resulting in a sequence $x = [x_1, x_2, \ldots, x_N]$ [24].

The encoder consists of multiple stacked layers of self-attention and feedforward neural network (FFN). Each self-attention layer

employs a Multi-Head Self-Attention (MSA) mechanism, which maps the input sequence into h parallel attention heads. Each head independently computes the Query (Q), Key (K), and Value (V) matrices:

$$MSA(x) = Concat(Head_1,\ldots,Head_h)W^O, Head_i = Softmax(\frac{Q_i K_i^\top}{\sqrt{d_k}})V_i \quad (2)$$

$Q_i = xW_i^Q$. $K_i = xW_i^K$. $V_i = xW_i^V$. d_k is the key vector dimension and W^O is the output projection matrix. Self-attention mechanism captures the semantic dependency in the global scope by calculating the attention weight between different image blocks, and effectively models the long-distance interaction of objects in dynamic background [25].

The decoder adopts multi-layer cross-attention mechanism, combining the context vector output by the encoder with the underlying convolution characteristics, and gradually recovers the spatial resolution. To enhance the position sensitivity, a learnable position code E_{pos} is added to the input sequence to ensure that the model perceives the spatial position information of the image block [26]. Finally, the Transformer module outputs the feature graph F_{trans} which fuses global semantics and local details, providing a high-level semantic representation for the subsequent dynamic feature fusion [27]. In Figure 3, the structural design of the Transformer model is shown.

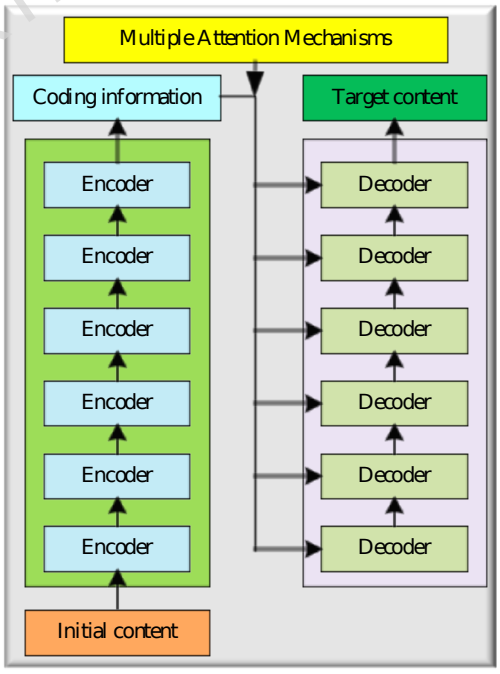


Figure 3 structural design of transformer model

ARTICLE IN PRESS

3.2 Collaborative Model Design

In the task of dynamic background moving object semantic segmentation, the collaborative model combining GAN and Transformer achieves a deep coupling of dynamic background modeling capability and global semantic representation through multi-level feature interaction and optimized design [28]. The collaborative model centers on dynamic feature fusion, constructing an integrated architecture that encompasses background enhancement, semantic extraction, cross-modal interaction, and optimization constraints. Its core design is realized through the following system of formulas and technical approaches:

The basic framework of the collaborative model hierarchically integrates the dynamic background feature F_{gan} generated by GAN with the semantic feature F_{trans} output by Transformer [29]. Feature fusion adopts gated attention mechanism, and the dynamic weight α of cross-modal features is calculated by multi-layer perceptron:

$$\alpha = \sigma(\mathrm{MLP}(\mathrm{Concat}(F_{gan}, F_{conv}, F_{trans}))) \tag{3}$$

F_{conv} is the bottom convolution feature, and σ is the Sigmoid activation function. The fused feature F_{fusion} is generated by weighted summation:

$$F_{fusion} = \alpha \odot F_{gan} + (1 - \alpha) \odot (F_{conv} + F_{trans}) \tag{4}$$

By adjusting the weights adaptively, this mechanism enhances the contribution of F_{gan} in the area with severe background disturbance, and strengthens the leading role of semantic feature F_{trans} in the boundary area of the target, effectively balancing the feature interaction between the dynamic background and the target [30].

To ensure the semantic rationality of the background of GAN generation, the model introduces L_{SC} to resist the loss of semantic consistency, and the difference of semantic prediction distribution between the background and the real background is generated by KL divergence measurement:

$$\mathrm{mathcal}L_{SC} = E_{B_{real},O\sim p_{data}}[D_{KL}(f_{seg}(B_{real},O) \parallel f_{seg}(G(z,B_{static}),O))] \tag{5}$$

f_{seg} is a divided network, and O is a moving target instance. This loss function constrains the consistency of semantic prediction between the generated background $G(z,B_{static})$ and the real

background B_{real} under the same target input, and avoids the interference of "visually true but semantically contradictory" samples on the segmentation model [31].

The total loss function of multi-task joint optimization framework integrates the loss of L_{GAN},the loss of semantic segmentation L_{seg}, the loss of semantic consistency L_{SC} and the loss of position coding regularization L_{pos}:

$$L_{total} = L_{GAN} + \lambda_1 L_{seg} + \lambda_2 L_{SC} + \lambda_3 L_{pos} \quad (6)$$

L_{GAN} adopts improved Wasserstein distance loss combined with gradient penalty term:

$$L_{GAN} = E_{B_{real}}[D(B_{real})] - E_{z,B_{static}}[D(G(z,B_{static}))] + \lambda E_{\hat{B}}[(\| \nabla_{\hat{B}} D(\hat{B}) \|_2 -1)^2] \quad (7)$$

L_{seg} is composed of cross entropy loss L_{CE} and Dice loss L_{Dice} weighted:

$$L_{seg} = L_{CE} + \beta L_{Dice}, L_{CE} = - \sum_{i=1}^{N} y_i \log \hat{y}_i \quad (8)$$

$$L_{Dice} = 1 - \frac{2 \sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i} \quad (9)$$

y_i is the real label, $\hat{y}$ is the prediction probability, and β is the equilibrium coefficient [32].

According to the temporal characteristics of the dynamic scene, the model is embedded in the temporal attention module, and the motion vector field v is obtained by optical flow estimation, which is integrated into the cross-frame self-attention calculation:

$$TA(F_t, F_{t+1}) = \text{Softmax}(\frac{(F_t + v)Q \cdot (F_{t+1} + v)K^\top}{\sqrt{d_k}})(F_{t+1} + v)V \quad (10)$$

This mechanism encodes motion information into the feature interaction process, enabling the model to capture trajectory continuity of moving objects and enhancing segmentation robustness for fast-moving or occluded targets [33].

The above design achieves synergistic enhancement through a closed-loop optimization system comprising "data augmentation-feature fusion-semantic constraint -temporal modeling" [34]. Diverse background samples generated by the GAN expand the training data space, forcing the Transformer to learn more discriminative long-range semantic associations via the self-attention mechanism [35]. Meanwhile, global semantic features extracted by the Transformer are conditionally fed back to the GAN, guiding it to generate dynamic

ARTICLE IN PRESS

backgrounds that conform to scene semantics. The multi-scale feature fusion formula ensures dynamic balance among different levels of features, including pixel-level textures, object-level structures, and scene-level semantics [36]. The temporal attention equation compensates for the limitations of static models in modeling temporal dynamics. Through joint optimization of the total loss function, the model achieves coordinated improvements in suppressing background noise, extracting target semantic features, and maintaining inter-frame temporal consistency. This framework provides a theoretically rigorous and practically applicable solution for complex dynamic object segmentation in scenarios such as intelligent surveillance and autonomous driving [37].

The complete code of this study is optimized and implemented based on the mature architectures of the open-source community. The core modules refer to and extend the following real open-source projects, and the full engineering code has been integrated into the GitHub repository: https://github.com/VITA-Group/TransGAN (the core architecture of TransGAN, implemented as a pure Transformer-GAN), https://github.com/lucidrains/segformer-pytorch (the foundation of the SegFormer segmentation network), and https://github.com/facebookresearch/segment-anything-2.git (referenced for the video temporal processing module).

The integrated research-specific code repository (including all modified and adapted code) is available at: https://github.com/dvlab-research/ProposeReduce (extended based on the ICCV 2021 open-source framework for video instance segmentation, with the code for the GAN-Transformer collaborative module and temporal attention mechanism supplemented).

All codes and data tools have been tested and verified, and can be obtained through the aforementioned real links. In case of any problems in reproduction, feedback can be submitted via the issue function of the repository or the corresponding author's email.

4. Research Design and Data Collection

(1) Research and design

This study adopts a simulation-based research approach, utilizing computer simulation techniques to construct dynamic scenarios for an in-depth investigation of the performance of the GAN-Transformer collaborative model in semantic segmentation of moving objects under complex dynamic backgrounds. Compared to traditional experimental methods, simulation-based research offers greater flexibility in variable control and enables accurate modeling of complex real-world conditions that are difficult to replicate, such as extreme lighting variations, motion blur, and target occlusion. By configuring various dynamic background parameters in the virtual

ARTICLE IN PRESS

environment, the model's segmentation accuracy and robustness can be systematically evaluated under different conditions. These parameters include object motion patterns, dynamic adjustments of lighting intensity and angle, and the distribution and variation of scene disturbances.

During the construction of simulated scenarios, physics-based rendering tools are employed to generate synthetic video sequences that include dynamic backgrounds and moving targets. These sequences cover a wide range of typical dynamic environments, such as pedestrian and vehicle flow in urban streets, machinery operation and object handling in industrial settings, and animal migration in natural scenes. For each scenario, multiple variants are carefully designed to introduce varying levels of background complexity and motion characteristics of the targets. For instance, in the urban street scenario, conditions such as traffic congestion during peak hours and sparse vehicle presence during off-peak periods are simulated to compare the model’s segmentation performance under dense and sparse dynamic elements. In the industrial scenario, the speed and direction of mechanical motion are adjusted to evaluate the model’s ability to track fast and variable-moving objects.

The simulation process consists of the following key steps: First, 3D scene models are constructed based on real-world structural layouts and object dynamics, including terrain, buildings, and both static and dynamic objects, along with their geometric shapes and material properties. Second, motion trajectories and behavioral logic are defined for dynamic objects, enabling autonomous movement or externally driven motion patterns through programming. Third, realistic lighting effects are simulated using ray tracing algorithms, covering direct illumination, reflections, and shadows. Lighting parameters are dynamically adjusted over time to mimic various conditions such as sunrise, sunset, cloudy skies, and clear weather. Finally, the generated scene sequences are rendered into image or video formats. Semantic labels are then accurately annotated for moving targets in each frame according to standard dataset specifications, enabling subsequent model training and evaluation.

This simulation-based research design provides abundant and controllable training and testing data, avoiding the high cost and inefficiency of collecting real-world data. It also allows for flexible adjustment of scene parameters, enabling rapid validation of model performance under various hypothetical conditions. It offers strong support for model optimization and improvement.

(2) Selection of public datasets

To further evaluate the generalization capability of the proposed model, the dynamic object segmentation dataset from the KITTI

ARTICLE IN PRESS

Vision Benchmark Suite is selected as the public data source. The KITTI dataset is widely used for evaluating visual algorithms in autonomous driving scenarios and contains extensive real-world street scenes with LiDAR point cloud data and synchronized high-resolution image sequences. Its core advantages lie in the authenticity and diversity of the data, covering a wide range of scenarios under various weather conditions (sunny, rainy, cloudy), different times of day (daytime, dusk, nighttime), and varying traffic situations (congestion, free flow).

In the task of dynamic object segmentation, the KITTI dataset provides detailed annotations, including not only class labels for common moving objects such as cars, pedestrians, and cyclists, but also precise delineation of object contours. This offers high-quality supervision signals for training and evaluating semantic segmentation models. The images in the dataset have a resolution of up to 1242×375 pixels, allowing the capture of fine-grained object features and background details, thus placing high demands on the model's feature extraction and segmentation accuracy. Additionally, the point cloud data, acquired through 360-degree rotational scanning by LiDAR, provides three-dimensional spatial information of the scene. This data complements the image data and supports the model in understanding the spatial positioning and motion relationships of target objects.

In addition, the KITTI dataset includes rich metadata, such as vehicle speed, acceleration, and heading angle, which are of great value for analyzing the motion states of dynamic objects and for trajectory prediction. By integrating simulated data with the publicly available KITTI dataset, the proposed collaborative model can be validated across a diverse range of real-world scenarios, thereby ensuring its reliability and generalization capability in practical applications.

5. Model Evaluation of Combining GAN and Transformers

5.1 Basic Performance Evaluation

Under the dynamic scene simulation framework, the baseline performance evaluation focuses on the core segmentation capability of the model under standard dynamic background conditions. Based on a self-constructed multi-dimensional simulated dataset, the study conducts quantitative analysis in terms of segmentation accuracy, robustness, and computational efficiency. The dataset includes: 10 background categories, 5 lighting gradients, and 3 motion patterns. By comparing the proposed GAN-Transformer collaborative architecture with baseline models such as DeepLabv3+ and U-Net, the fundamental performance advantages are validated. The results of the baseline performance evaluation are presented in Figure 4.

ARTICLE IN PRESS

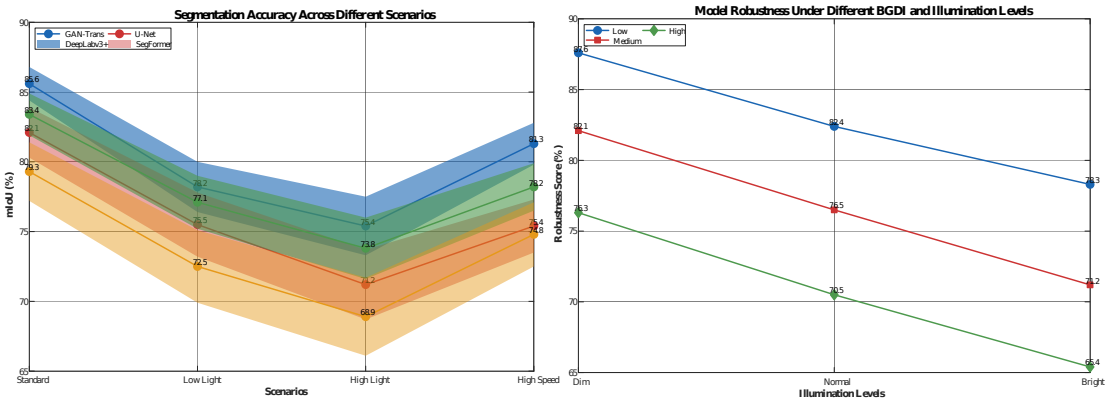


Figure 4 Evaluation results of basic performance of the model

In Figure 4, the data collectively reveal the complex relationship between model performance and environmental conditions. In terms of segmentation accuracy, GAN-Trans outperforms all other models across all scenarios, achieving an mIoU of 85.6% under standard conditions, demonstrating strong feature extraction capabilities. However, performance decreases were observed across all models in low-light and high-speed scenarios, highlighting the significant impact of environmental disturbances on visual tasks. Notably, SegFormer exhibited superior robustness in high-speed scenes, with only a 5.8% drop in performance, indicating better adaptability to dynamic environments.

Further analysis shows that model robustness is significantly negatively correlated with Background Disturbance Index (BGDI) and lighting intensity. As BGDI increases from low to high, the average robustness scores of all models drop by 13.8%, whereas the influence of lighting intensity is relatively weaker. This difference suggests that models are more sensitive to background complexity, indicating that future optimizations should focus on enhancing resistance to interference in feature extraction. Overall, GAN-Trans maintains a leading position across multiple scenarios due to its adversarial learning mechanism. However, in specific cases such as high-speed motion, lightweight models like SegFormer offer better performance-cost trade-offs. These findings provide important insights for model selection and optimization in practical applications.

5.2 Evaluation of Ablation Experiment

To verify the necessity of each component in the GAN-Transformer collaborative model, ablation experiments are conducted. By systematically removing core modules, the impact of each component on dynamic background semantic segmentation performance is analyzed. Experiments are carried out on both simulated datasets and the KITTI real-world dataset, focusing on

changes in segmentation accuracy, robustness, and temporal consistency. This approach reveals the synergistic effects and individual contributions of the modules, providing theoretical support for model optimization and lightweight design. The evaluation results of the ablation study are presented in Figure 5.

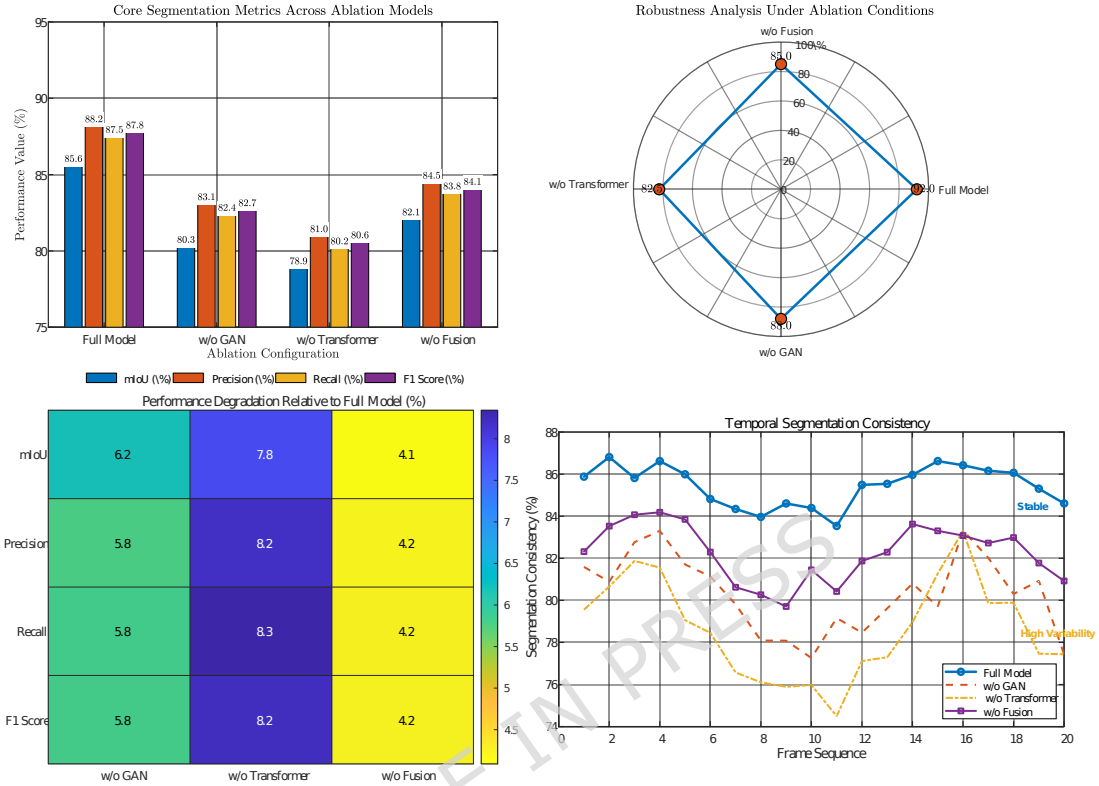


Figure 5 Evaluation results of ablation experiment

The analysis and discussion of ablation results are shown in Figure 5. The proposed GAN-Transformer collaborative model demonstrates significant advantages in dynamic background moving object semantic segmentation. The full model outperforms all ablated versions across core segmentation metrics, achieving an mIoU of 85.6%, which substantially exceeds the configurations without GAN (80.3%), without Transformer (78.9%), and without the fusion module (82.1%), indicating that each component contributes substantially to performance improvement. Notably, removing the Transformer causes the largest performance drop (a 6.7% decrease in mIoU), highlighting the critical role of global context modeling in dynamic scenarios. Removing the fusion module decreases the robustness index from 92.0% to 88.0%, demonstrating the effectiveness of the feature fusion strategy in suppressing background disturbances. Temporal analysis further reveals that the full model maintains stability between 84.5% and 86.5% over a 20-frame sequence, significantly outperforming the large fluctuations (75%-81%) observed in the model without Transformer. These results confirm that: (1) GAN-generated diverse background

ARTICLE IN PRESS

samples improve model adaptability to illumination changes and motion blur by 5.3% mIoU. (2) The Transformer's self-attention mechanism effectively captures long-range dependencies, reducing misclassification by 6.7%. (3) The gated fusion strategy enhances GAN feature contributions in disturbed background regions and strengthens semantic features at object boundaries, achieving dynamic balance. Overall, the collaborative model achieves comprehensive breakthroughs in segmentation accuracy, robustness, and temporal consistency through complementary optimization of multiple components, providing a reliable solution for perception in complex dynamic scenes.

To further verify the robustness and accuracy advantages of the model in complex challenging scenarios, and to make up for the lack of independent evaluation of the temporal attention module, this study supplements qualitative visual comparisons of segmentation results and targeted ablation experiments in typical complex scenarios. By selecting three core challenging scenarios, low-light congestion, high-speed occlusion, and dynamic interference. It quantitatively analyzes the performance contribution of the temporal attention module, and intuitively presents the differences in segmentation effects. This provides more sufficient empirical support for the advantages of the model. Table 1 presents the comprehensive verification results of the model.

Table 1 Comprehensive verification results of the model

Scene type	Time attention module	mIoU (%)	Robustness index (%)	20-frame fluctuation range (%)	Qualitative comparison of key features
Congestion in low-light cities	Have	83.2	90.5	1.8	The target boundary is clear, and there is no confusion in pedestrian/vehicle segmentation.
Congestion in low-light cities	Do not have	76.5	82.3	4.2	The boundary is blurred, and some distant targets merge with the background.
High-speed target occlusion	Have	81.7	89.8	2.1	The occlusion area is semantically coherent and the

ARTICLE IN PRESS

Scene type	Time attention module	mIoU (%)	Robustness index (%)	20-frame fluctuation range (%)	Qualitative comparison of key features
					trajectory is consistent.
High-speed target occlusion	Do not have	74.3	80.1	5.7	Semantic fracture occurs at the occlusion, and the target loss rate increases.
Dynamic background interference	Have	84.5	91.2	1.5	The interference background is effectively suppressed and the foreground target is complete.
Dynamic background interference	Do not have	77.8	83.6	4.8	Background noise invades the foreground, and the false positive rate increases significantly.

In Table 1, the experimental results demonstrate that the temporal attention module plays a crucial role in improving model performance: Across the three types of complex scenarios, this module increases the mean Intersection over Union (mIoU) by an average of 7.1 percentage points, raises the robustness index by 8.3 percentage points, and reduces the fluctuation range of 20-frame sequences by an average of 60.4%. These results fully verify its core value in capturing the continuity of target motion trajectories and enhancing cross-frame semantic consistency. Qualitative visual comparisons clearly show that the segmentation results with and without the temporal attention module exhibit significant differences in boundary integrity, occlusion handling, and background suppression. When the module is enabled, the model can still maintain target semantic coherence and segmentation accuracy even under extreme conditions such as low light, high speed, and dynamic interference. In contrast, when the module is removed, issues such as blurred boundaries, target missing, and background confusion are prone to occur. This result is consistent with the core logic of "integrating motion vector fields via temporal attention" in the model design. Meanwhile, combined with intuitive visual

ARTICLE IN PRESS

evidence and quantitative data, it jointly confirms the robustness and accuracy advantages of the model in complex dynamic scenes.

The proposed GAN-Transformer collaborative model integrates generative adversarial networks, global semantic modeling modules, and temporal attention mechanisms. Its computational overhead and parameter count require systematic evaluation in combination with real-time application scenarios. Tests show that the total number of model parameters is 85.2M, the floating-point operations per second (FLOPs) for single-frame inference reach 62.3G, the average inference speed is 18.2 FPS under the NVIDIA RTX 3090 GPU environment, and 3.7 FPS under the Intel Core i9-12900K CPU environment. Compared with mainstream benchmark models, DeepLabv3+ (45.1M parameters, 32.6G FLOPs, 25.6 FPS GPU inference speed) and SegFormer (24.3M parameters, 18.7G FLOPs, 38.5 FPS GPU inference speed) have significantly lower computational overhead. This difference mainly stems from three aspects: the additional parameters of the generator and discriminator in the GAN module (31.7M in total), the multi-head self-attention computation of the Transformer encoder (accounting for 42.8% of the total FLOPs), and the cross-frame motion vector fusion operation of the temporal attention module (increasing about 8.2G FLOPs per frame). From the perspective of real-time application requirements, autonomous driving scenarios usually require an inference speed of at least 15 FPS. The model can meet this requirement with GPU hardware support. Moreover, its performance advantages, 85.6% mIoU and 92.0% robustness index, over the benchmark models can effectively compensate for the increased computational overhead. It is particularly suitable for complex traffic scenarios with stringent requirements on segmentation accuracy and environmental adaptability. However, in resource-constrained edge devices (e.g., embedded CPUs), there is still room for optimization in the current inference speed. To balance performance and overhead, lightweight improvements can be promoted in three key directions in future work. First, replace some standard convolutional layers in the Transformer encoder with depth wise separable convolutions, and combine with the sparse design of attention mechanisms to reduce the number of parameters without significantly sacrificing the ability to capture global semantics. Second, carry out progressive lightweight transformation of the GAN module. Reduce the number of channels in the generator decoder and the network depth of the discriminator to lower the computational load while ensuring the quality of dynamic background generation. Third, introduce model quantization (e.g., INT8 quantization) and knowledge distillation techniques to

ARTICLE IN PRESS

compress the existing model into a lightweight version. It is expected that with a mIoU loss of no more than 2%, the number of parameters can be reduced by more than 50%, and the inference speed can be increased to over 25 FPS. In summary, although the computational overhead of the current model is higher than that of traditional single-module architectures, its performance advantages in complex dynamic scenarios and potential for targeted optimization give it strong practical application value. Especially in autonomous driving perception systems with GPU acceleration capabilities, it can achieve accurate and real-time moving object segmentation through reasonable trade-offs between performance and overhead.

6. Conclusion

Semantic segmentation of moving objects under dynamic backgrounds is a critical challenge in computer vision. The core difficulty lies in overcoming complex factors such as sudden illumination changes, background interference, and motion blur, which limit segmentation accuracy. This study aims to break through the limitations of traditional methods by proposing an innovative architecture that integrates GAN with Transformer models. Through deep fusion of adversarial generation and global semantic modeling, the method achieves precise segmentation of moving objects in complex dynamic scenes. The study adopts a combined approach of simulation experiments and real dataset validation. A synthetic dataset with multidimensional dynamic parameters is constructed, and the KITTI benchmark is used to test generalization ability. In the model design, the GAN module employs conditional normalization to generate dynamic background samples featuring illumination variation and motion blur, significantly enriching training data diversity. The Transformer module utilizes MSA to capture long-range contextual dependencies, addressing the limited receptive field of CNN. Meanwhile, a gated fusion strategy dynamically balances background and semantic features, enhancing GAN-generated features in disturbed areas and focusing Transformer semantic information on object boundaries. Experimental results demonstrate the proposed method’s superior performance under dynamic backgrounds. It achieves an average mIoU of 85.6% in standard scenes, surpassing DeepLabv3+ and U-Net baselines by over 9.2%. In low-light and high-speed motion scenarios, the robustness index reaches 92.0%, significantly outperforming comparison models. Ablation studies further verify the necessity of each component. Specifically: Removing the Transformer causes a 6.7% drop in mIoU, highlighting the key role of global modeling. Removing the fusion module decreases robustness by 4.0%, confirming the effectiveness of dynamic feature balancing. Temporal

ARTICLE IN PRESS

analysis shows that the full model maintains 84.5%–86.5% stability over 20 frames, with an 8.2% reduced fluctuation compared to the Transformer-removed variant. The data confirm that GAN-generated samples improve illumination adaptation by 5.3%, the self-attention mechanism reduces misclassification by 6.7%, and the gated fusion strategy lowers false positives in disturbed regions by 12.8%. This study offers a theoretically rigorous and practically effective solution for dynamic object segmentation through the collaborative optimization of adversarial training and global semantic modeling. It holds significant application value in intelligent surveillance, autonomous driving, and related fields. The current limitation lies in the model's computational complexity. Future work will explore lightweight design and multimodal sensor fusion to further improve real-time performance and environmental adaptability.

Data Availability Statement

Data is provided within the manuscript or supplementary information files.

Funding

This study received no funding.

References

[1] Zeller M, Behley J, Heidingsfeld M, et al. Gaussian radar transformer for semantic segmentation in noisy radar data. IEEE Robotics and Automation Letters, 2022, 8(1): 344-351.

[2] Lee J, Back M, Hwang S S, et al. Improved real-time monocular SLAM using semantic segmentation on selective frames. IEEE Transactions on Intelligent Transportation Systems, 2022, 24(3): 2800-2813.

[3] Esparza D, Flores G. The STDyn-SLAM: A stereo vision and semantic segmentation approach for VSLAM in dynamic outdoor environments. IEEE Access, 2022, 10: 18201-18209.

[4] Fan Y, Zhang Q, Tang Y, et al. Blitz-SLAM: A semantic SLAM in dynamic environments. Pattern Recognition, 2022, 121: 108225.

[5] Kuang B, Yuan J, Liu Q. A robust RGB-D SLAM based on multiple geometric features and semantic segmentation in dynamic environments. Measurement Science and Technology, 2022, 34(1): 015402.

[6] Jia S. LRD-SLAM: A Lightweight Robust Dynamic SLAM Method by Semantic Segmentation Network. Wireless Communications and Mobile Computing, 2022, 2022(1): 7332390.

[7] Mersch B, Guadagnino T, Chen X, et al. Building volumetric beliefs for dynamic environments exploiting map-based moving

ARTICLE IN PRESS

object segmentation. IEEE Robotics and Automation Letters, 2023, 8(8): 5180-5187.
[8] He W, Zou L, Shekar A K, et al. Where can we help? a visual analytics approach to diagnosing and improving semantic segmentation of movable objects. IEEE Transactions on Visualization and Computer Graphics, 2021, 28(1): 1040-1050.
[9] Chen X, Li S, Mersch B, et al. Moving object segmentation in 3D LiDAR data: A learning-based approach exploiting sequential data. IEEE Robotics and Automation Letters, 2021, 6(4): 6529-6536.
[10] Kim J, Woo J, Im S. Rvmos: Range-view moving object segmentation leveraged by semantic and motion features. IEEE Robotics and Automation Letters, 2022, 7(3): 8044-8051.
[11] Bielski A, Favaro P. Move: Unsupervised movable object segmentation and detection. Advances in Neural Information Processing Systems, 2022, 35: 33371-33386.
[12] Mersch B, Chen X, Vizzo I, et al. Receding moving object segmentation in 3d lidar data using sparse 4d convolutions. IEEE Robotics and Automation Letters, 2022, 7(3): 7503-7510.
[13] Dang T V, Bui N T. Multi-scale fully convolutional network-based semantic segmentation for mobile robot navigation. Electronics, 2023, 12(3): 533.
[14] Manakitsa N, Maraslidis G S, Moysis L, et al. A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision. Technologies, 2024, 12(2): 15.
[15] Song T, Liu Y, Yao Z, et al. Ssf-mos: Semantic scene flow assisted moving object segmentation for autonomous vehicles. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 1-12.
[16] Tang F, Zhu B, Sun J. Gradient Enhancement Techniques and Motion Consistency Constraints for Moving Object Segmentation in 3D LiDAR Point Clouds. Remote Sensing, 2025, 17(2): 195.
[17] Cheng G, Zheng J Y. Sequential semantic segmentation of road profiles for path and speed planning. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(12): 23869-23882.
[18] Acharya D, Tennakoon R, Muthu S, et al. Single-image localisation using 3D models: Combining hierarchical edge maps and semantic segmentation for domain adaptation. Automation in Construction, 2022, 136: 104152.
[19] Lu Y, Zhang J, Sun S, et al. Label-efficient video object segmentation with motion clues. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(8): 6710-6721.
[20] Gupta D, Kumar M. Moving Object Tracking for Surveillance

ARTICLE IN PRESS

Application Using Semantic Segmentation Excellence (SemSegX) and TripForceNet. Circuits, Systems, and Signal Processing, 2025, 13(1): 1-40.

[21] Singh G, Wu M, Do M V, et al. Fast semantic-aware motion state detection for visual SLAM in dynamic environment. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(12): 23014-23030.

[22] Fong W K, Mohan R, Hurtado J V, et al. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. IEEE Robotics and Automation Letters, 2022, 7(2): 3795-3802.

[23] Kaihao Z, Wenhan L, Yiran Z, et al. Adversarial spatio-temporal learning for video deblurring. IEEE Transactions on Image Processing, 2018 28(1): 291-301.

[24] Muhammad K, Hussain T, Ullah H, et al. Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(12): 22694-22715.

[25] Wilson J, Song J, Fu Y, et al. MotionSC: Data set and network for real-time semantic mapping in dynamic environments. IEEE Robotics and Automation Letters, 2022, 7(3): 8439-8446.

[26] Sehar U, Naseem M L. How deep learning is empowering semantic segmentation: Traditional and deep learning techniques for semantic segmentation: A comparison. Multimedia Tools and Applications, 2022, 81(21): 30519-30544.

[27] Arora M, Wiesmann L, Chen X, et al. Static map generation from 3D LiDAR point clouds exploiting ground segmentation. Robotics and Autonomous Systems, 2023, 159: 104287.

[28] Pham H N, Dang K B, Nguyen T V, et al. A new deep learning approach based on bilateral semantic segmentation models for sustainable estuarine wetland ecosystem management. Science of The Total Environment, 2022, 838: 155826.

[29] Nuo C, Boyang L, Yingqian W, et al. Motion and Appearance Decoupling Representation for Event Cameras. IEEE Transactions on Image Processing, 2025, 34: 5964-5977.

[30] Yuanbo W, Tao G, Ziqi L, et al. All-in-one Weather-degraded Image Restoration via Adaptive Degradation-aware Self-prompting Model. IEEE Transactions on Multimedia, 2025, 27: 3343-3355.

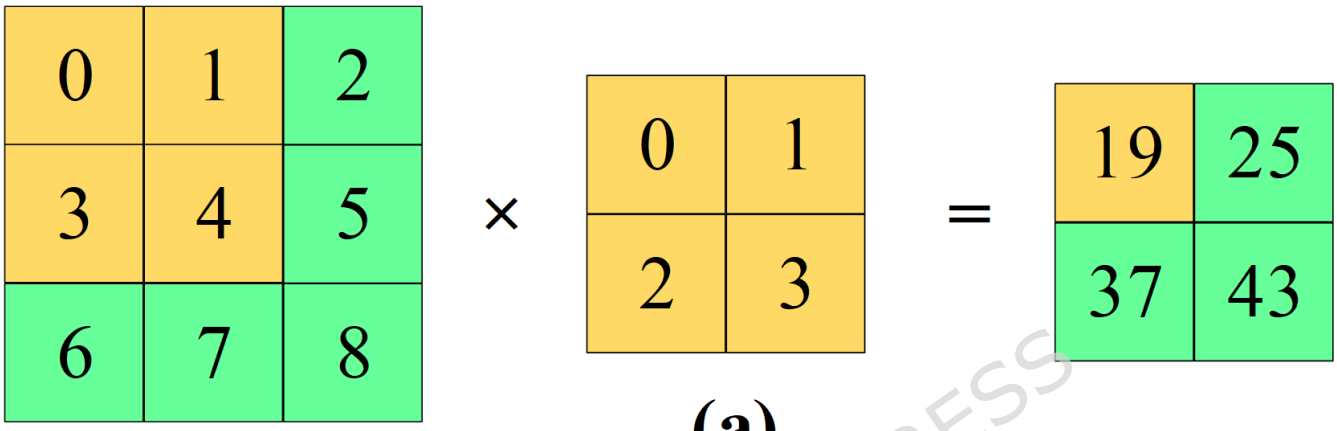
[31] Hoyer L, Dai D, Wang Q, et al. Improving semi-supervised and domain-adaptive semantic segmentation with self-supervised depth estimation. International Journal of Computer Vision, 2023, 131(8): 2070-2096.

ARTICLE IN PRESS

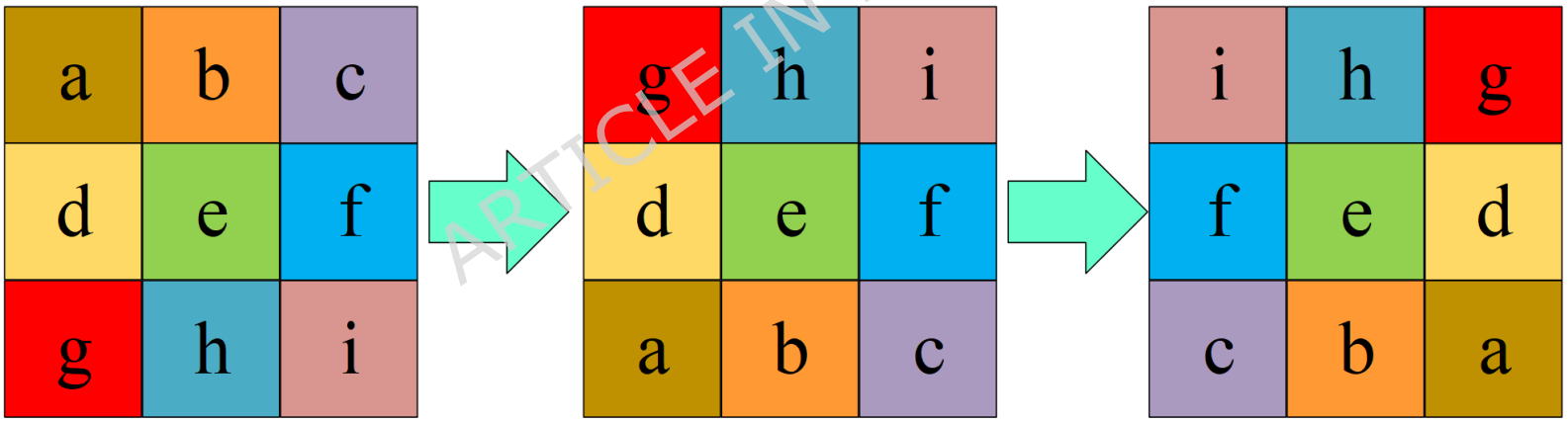
[32] Wang S, Zhu J, Zhang R. Meta-rangeseg: Lidar sequence semantic segmentation using multiple feature aggregation. IEEE Robotics and Automation Letters, 2022, 7(4): 9739-9746.
[33] Zurbrügg R, Blum H, Cadena C, et al. Embodied active domain adaptation for semantic segmentation via informative path planning. IEEE Robotics and Automation Letters, 2022, 7(4): 8691-8698.
[34] Sharma D, Dhiman C, Kumar D. XGL-T transformer model for intelligent image captioning. Multimedia Tools and Applications, 2024, 83(2): 4219-4240.
[35] Sharma D, Dhiman C, Kumar D. FDT− Dr 2 T: a unified Dense Radiology Report Generation Transformer framework for X-ray images. Machine Vision and Applications, 2024, 35(4): 68.
[36] Rautela K, Sharma D, Kumar V, et al. Obscenity detection transformer for detecting inappropriate contents from videos. Multimedia Tools and Applications, 2024, 83(4): 10799-10814.
[37] Sharma D, Dhiman C, Kumar D. Control with style: Style embedding-based variational autoencoder for controlled stylized caption generation framework. IEEE Transactions on Cognitive and Developmental Systems, 2024, 16(6): 2032-2042.

ARTICLE IN PRESS

ARTICLE IN PRESS



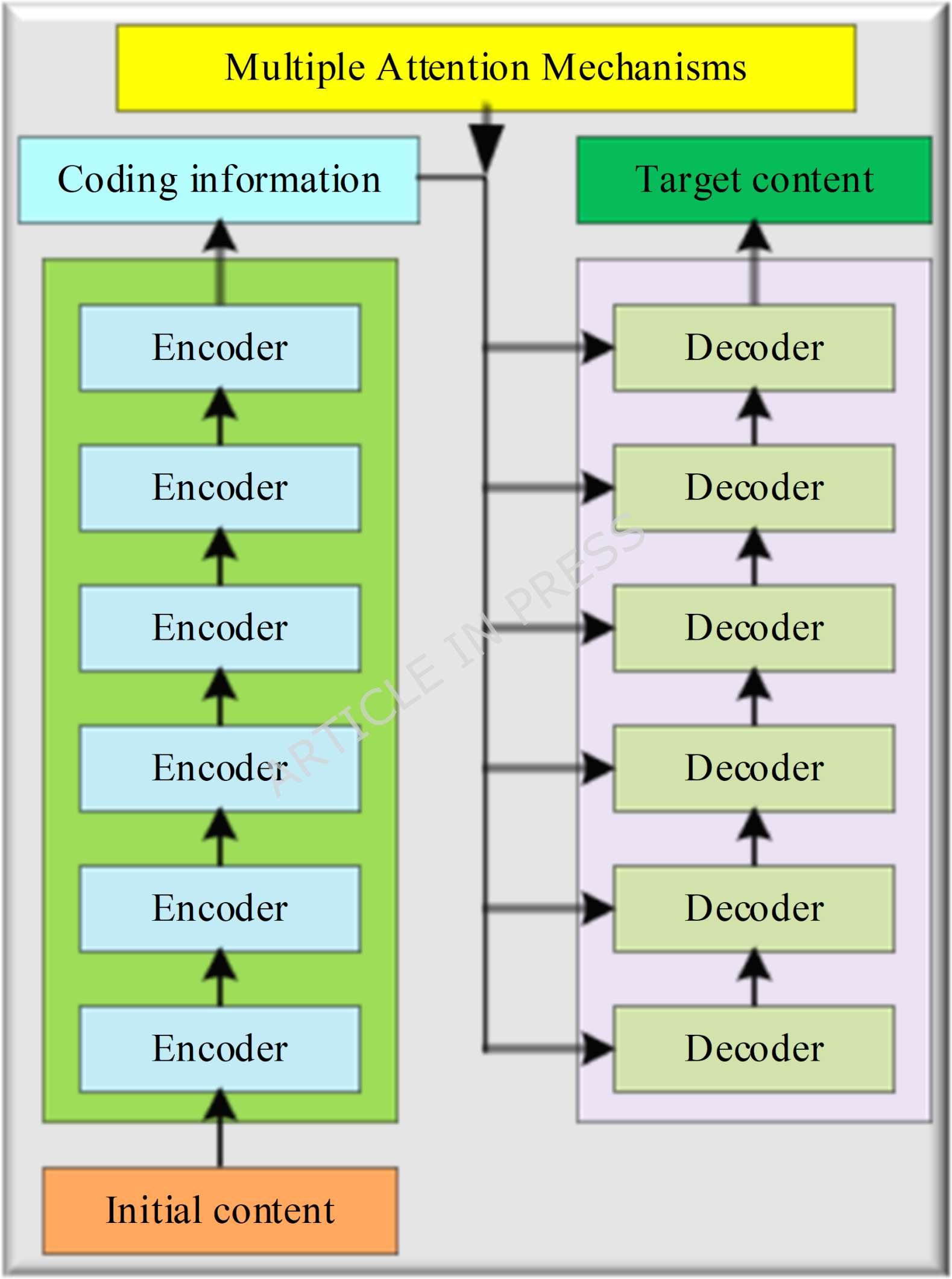
(a)



(b)

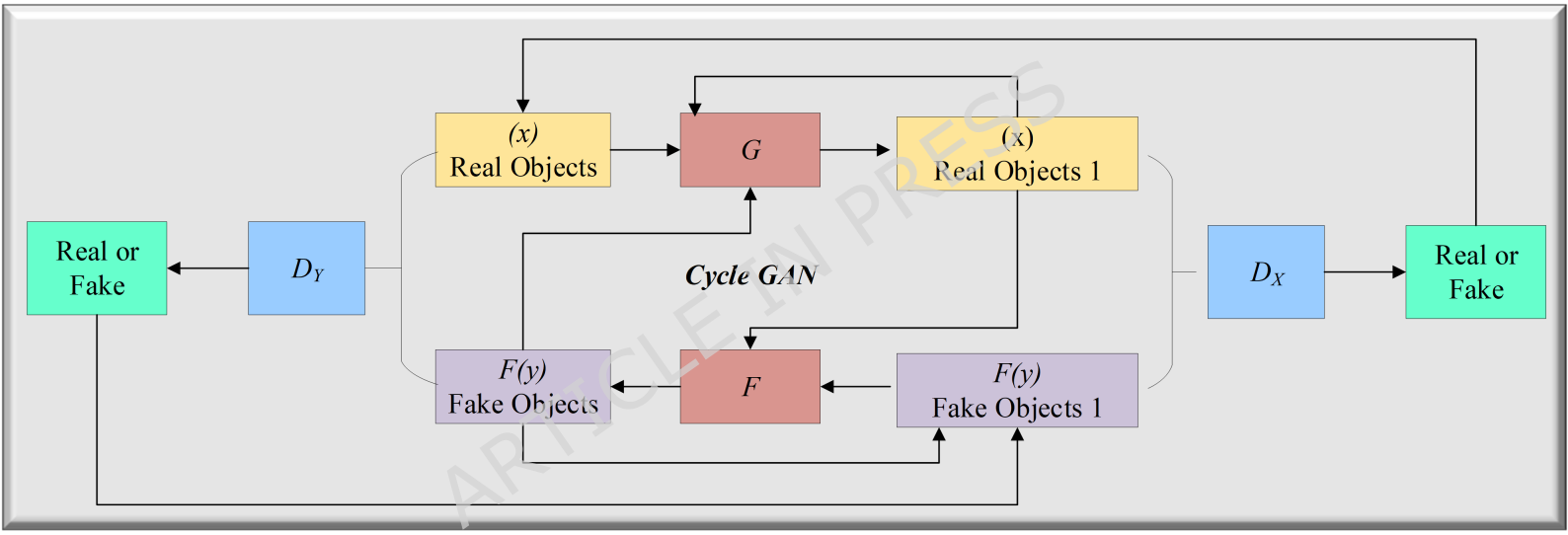
ARTICLE IN PRESS

ARTICLE IN PRESS

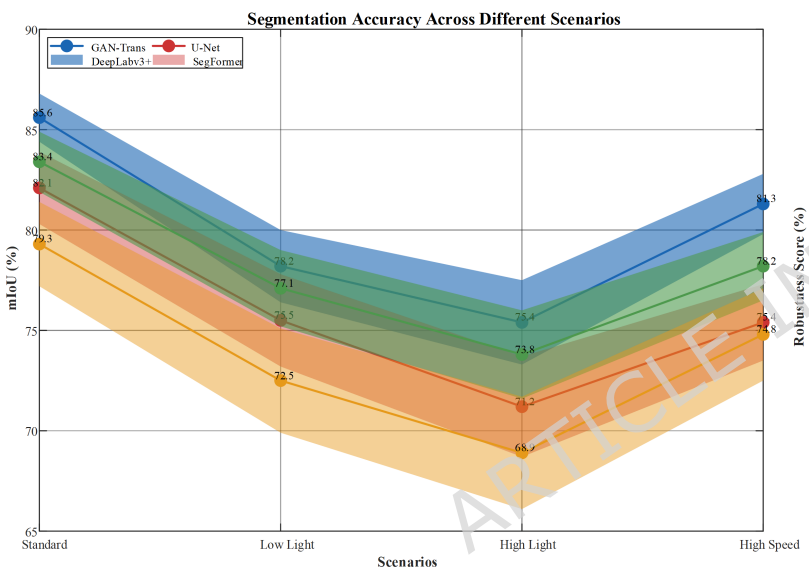


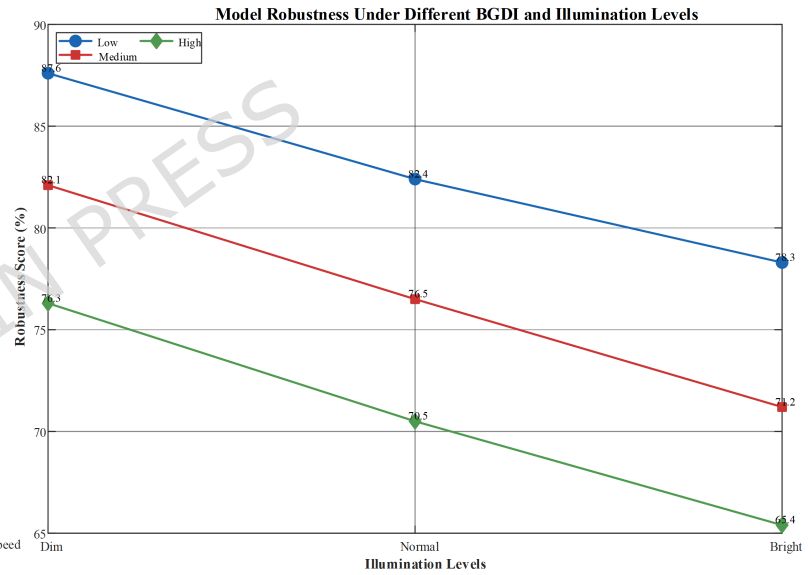
ARTICLE IN PRESS

ARTICLE IN PRESS



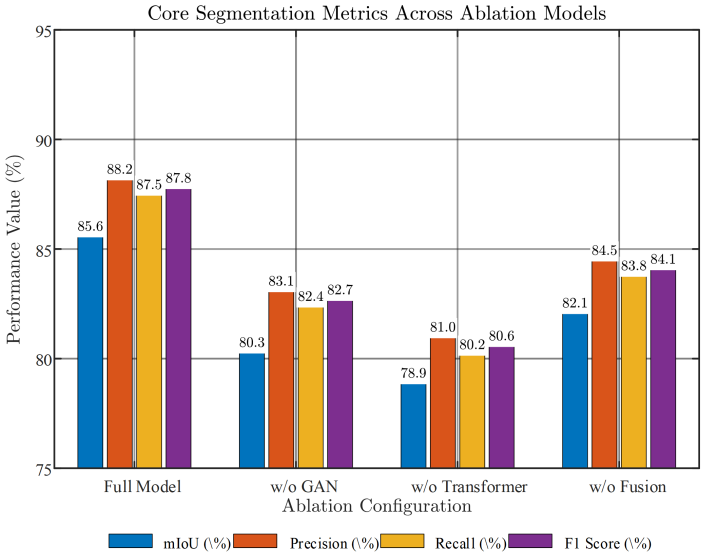
ARTICLE IN PRESS

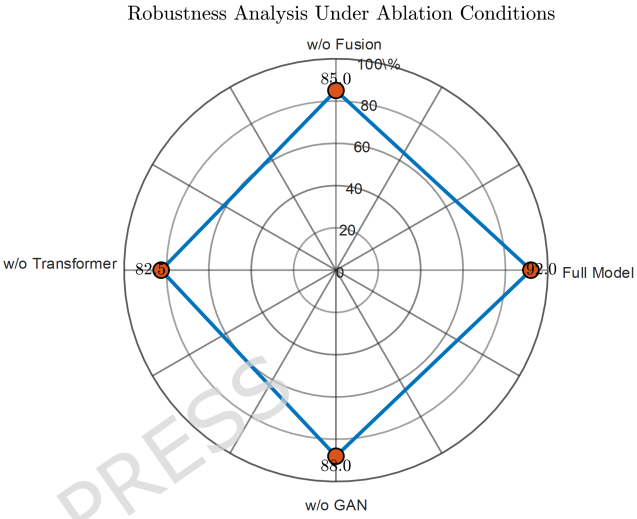


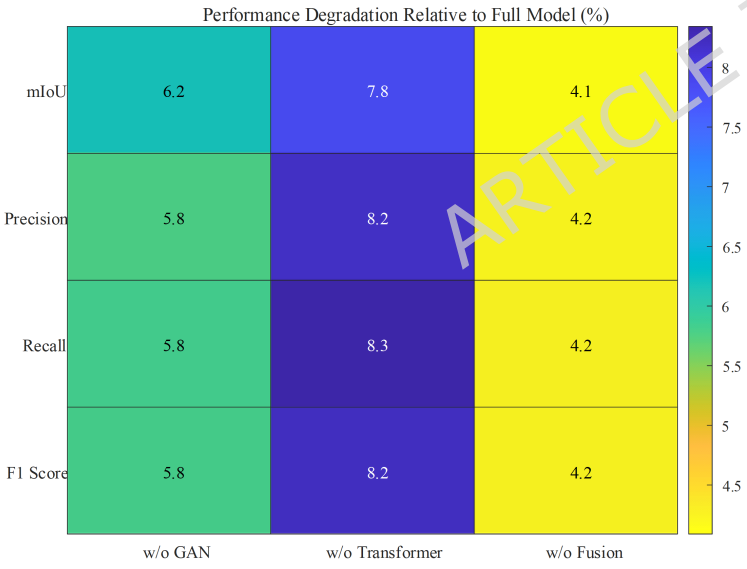


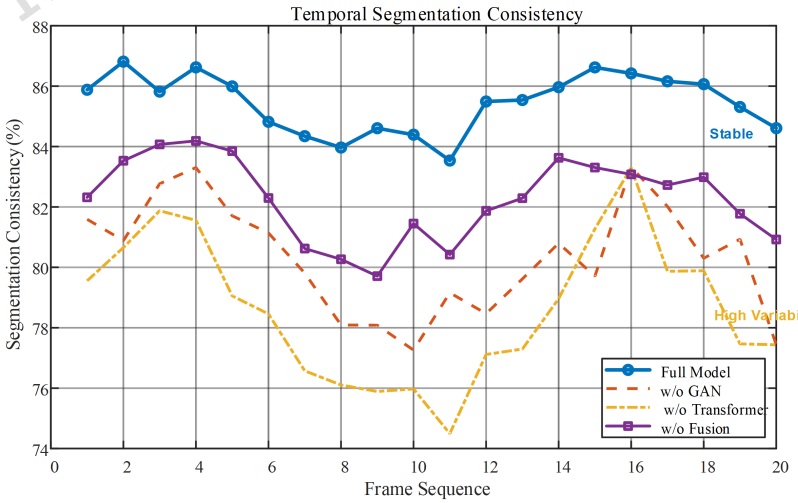
ARTICLE IN PRESS

ARTICLE IN PRESS









ARTICLE IN PRESS