Robust Instance Segmentation with Dynamic Gating Fusion and Adaptive Matching for Stable Training

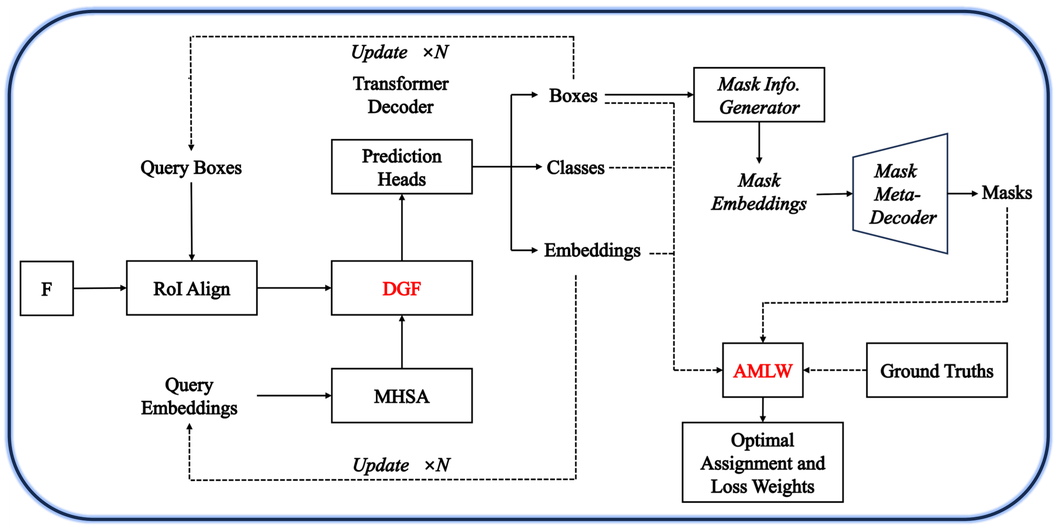
Zhimeng Jiao, Jinnan Zhang, Yutao Shi, Zheyu Liu, Xiaotian Yang, Zhenhai Li, Haitao Zhang, Xia Zhang

School of Electronic Engineering, Beijing University of Posts and Telecommimications, Beijing, China

Introduction

Instance segmentation, combining object detection and semantic segmentation, has been advanced by Transformer architectures. Notably, the masked embedding mechanism in Transformer improves accuracy by integrating spatial and embedding features. However, static fusion operations fail to adaptively combine spatial and embedding features, while fixed-cost Hungarian matching leads to sub-optimal box-mask alignment. To address these, we propose two lightweight modules: Dynamic Gating Fusion (DGF), which adaptively balances spatial and embedding features via learnable gates, and Adaptive Matching and Loss Weighting (AMLW), which dynamically adjusts matching costs and loss weights for better box-mask consistency.

Methods



The figure illustrates our instance segmentation framework built upon ISTR. Its core innovations are two modules: Dynamic Gating Fusion (DGF) adaptively merges spatial features and query embeddings to integrate fine-grained boundaries with high-level semantics, while Adaptive Matching and Loss Weighting (AMLW) dynamically optimizes matching costs and loss weights to enhance box-mask supervision consistency. The framework iteratively refines predictions via a Transformer decoder and generates final masks through a mask meta-decoder.

Conclusions

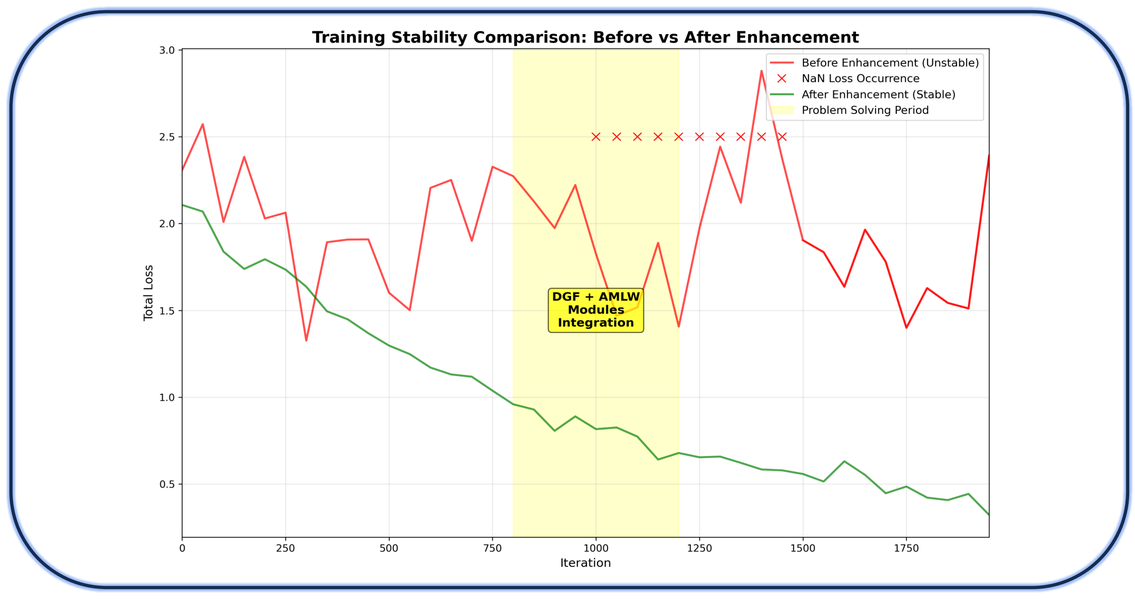
We propose two lightweight plug-and-play modules—DGF and AMLW—to address limitations in feature fusion and matching for transformer-based instance segmentation. DGF enhances feature representation and convergence via adaptive spatial-embedding fusion, while AMLW improves training consistency through dynamic matching and loss weighting. Together they mitigate instability and convergence issues in ISTR. The experimental results show that with the incorporation of DGF and AMLW, the training process exhibits smoother convergence curves, significantly reduces numerical instability, and lowers both the NaN rate and error counts, leading to a more stable and robust training process.

Future work

In future work, we plan to explore more advanced gating mechanisms—drawing inspiration from multimodal learning—to further enhance adaptive feature fusion. Additionally, we aim to extend the dynamic optimization strategy of AMLW to other matching-dependent vision tasks, seeking broader improvements in training consistency and generalization.

Results and discussion

The proposed DGF and AMLW modules complement each other at both the architectural and optimization levels, effectively improving the stability of the training process. As illustrated in figure, the enhanced framework achieves smoother convergence curves and more predictable updates compared with the baseline, avoiding oscillations and NaN errors while maintaining consistent convergence trajectories. These results confirm the synergistic roles of DGF and AMLW in significantly enhancing training stability.



Ablation results demonstrate that combining DGF and AMLW modules produces a notable synergistic effect. The joint configuration achieves optimal performance: a 99.31% loss reduction rate and a 0.58% NaN rate, significantly outperforming either module used alone. This confirms their strong complementarity in enhancing training effectiveness, where the synergy simultaneously improves both convergence efficiency and training stability.

Configuration	Loss Reduction(%)	Final Loss	NaN Count	NaN Rate(%)
DGF Only	84.41%	0.5286	355	0.87%
AMLW Only	79.44%	0.5726	426	1.04%
DGF+AMLW	99.31%	0.4405	237	0.58%