Enhancing Fine-grained Multi-modal Alignment via Adapters
A Parameter-Efficient Training Framework for Referring Image Segmentation

Zunnan Xu*, Jiaqi Huang*, Ting Liu, Yong Liu, Haonan Han, Kehong Yuan†, Xiu Li†
Tsinghua University

ICML International Conference On Machine Learning

Introduction

In this study, we address the challenge of efficient training in computer vision, particularly for multi-modal dense prediction tasks. We introduce DenseCrossAdapter, a parameter-efficient module that enhances feature propagation and cross-modal interaction in referring image segmentation.

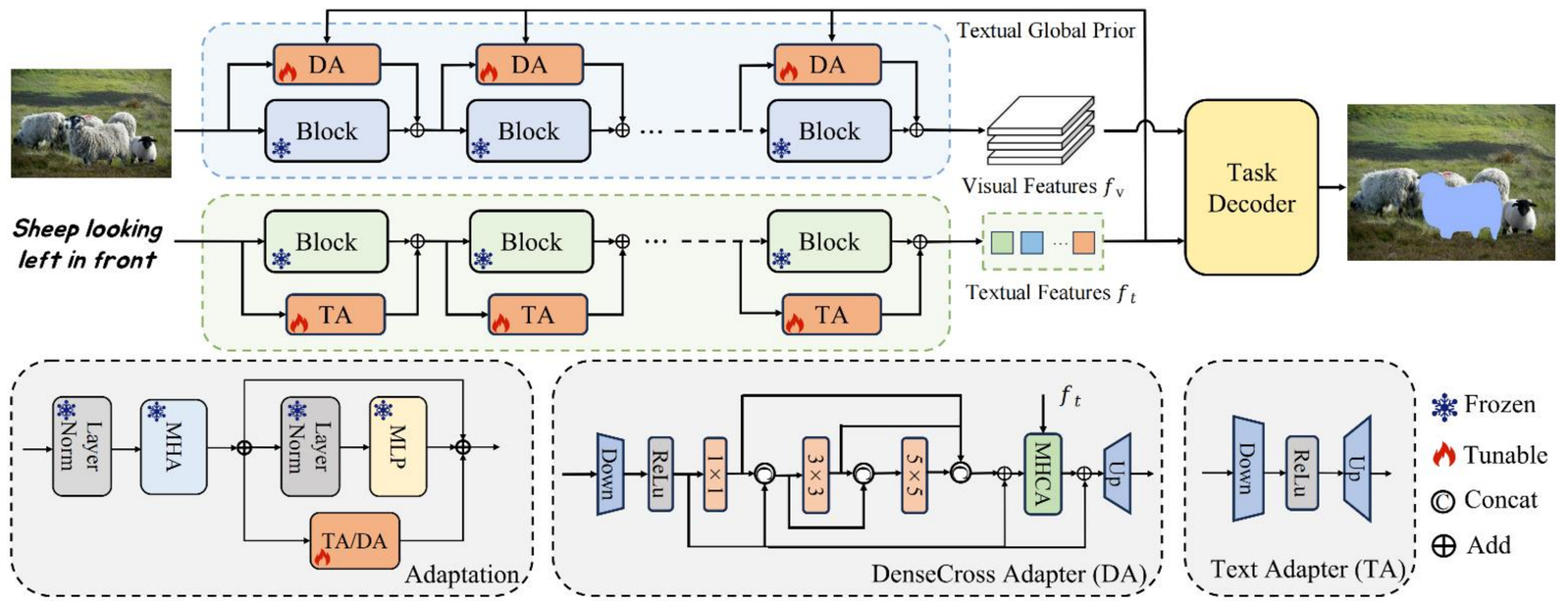
Contributions

(1) We apply the pre-trained model DINO in RIS tasks and propose an efficient training strategy for precise alignment without requiring complex design.

(2) We propose a novel DenseCrossAdapter that seamlessly integrates into a pre-trained backbone to enhance and interact with its intermediate features. This integration improves DINO's alignment with language and enhances its performance on dense prediction tasks.

(3) Experiments show that our method significantly outperforms state-of-the-art fully fine-tuned methods in image segmentation, with only 0.9% to 1.8% updates to the backbone parameters.

Overall Framework



Results on RefCOCO/RefCOCO+/G-Ref

(1) Our approach achieves state-of-the art performance with only 0.9% to 1.8% backbone parameter updates.

(2) Our method shows an IoU improvement of 2.82% over other parameter-efficient methods while using a comparable amount of fine-tuned parameters.

(3) Our method outperforms the state-of-the-art method that used full fine-tuning on mixed datasets, demonstrating good data scalability.

Scan me

Please scan the QR for our full paper.