A Survey on Open-Vocabulary Semantic Segmentation: Challenges, Methods and Future

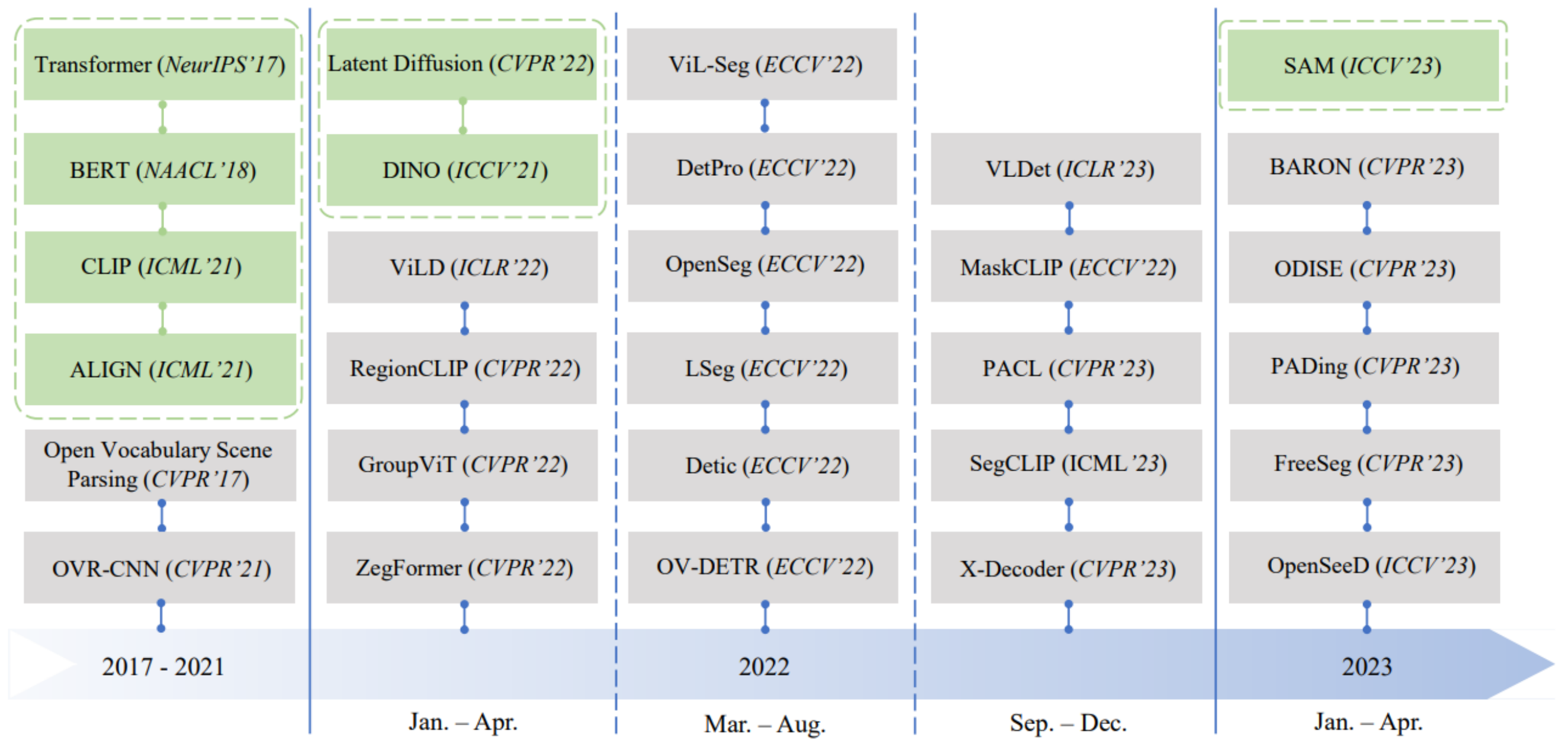
Guanlin Liu

中山大學
SUN YAT-SEN UNIVERSITY

Introduction and Background

➢ Introduction

Semantic segmentation aims to partition images into regions with semantic labels, but traditional methods are limited to a fixed set of categories. Open Vocabulary Semantic Segmentation (OVSS) overcomes this by enabling models to recognize and segment unseen categories using vision-language models like CLIP, and this review summarizes recent developments, methods, and future directions in OVSS.



➢ Preliminaries

I. Traditional Close-Set Semantic Segmentatio

The earliest work utilizing neural networks for semantic segmentation was proposed by Jonathan Long et al. in 2015. Their work was the first to introduce the conversion of traditional CNNs into Fully convolutional networks(FCN) for end-to-end pixel-level semantic segmentation, regarding as a cornerstone of modern semantic segmentation research, establishing the foundation for subsequent models such as U-Net and Deeplab.

II. Vision-Language Pre-Trained Models(VLPMS)

VLPMS represent a significant advancement in both computer vision and natural language processing in recent years.These models are trained on joint image and text data, learning to map images and texts into a common embedding space.

III. Transfer Learning and Self-Supervised Learning

Transfer learning and self-supervised learning are essential techniques for enhancing a model's generalization ability, enabling it to adapt to new categories and data distributions.

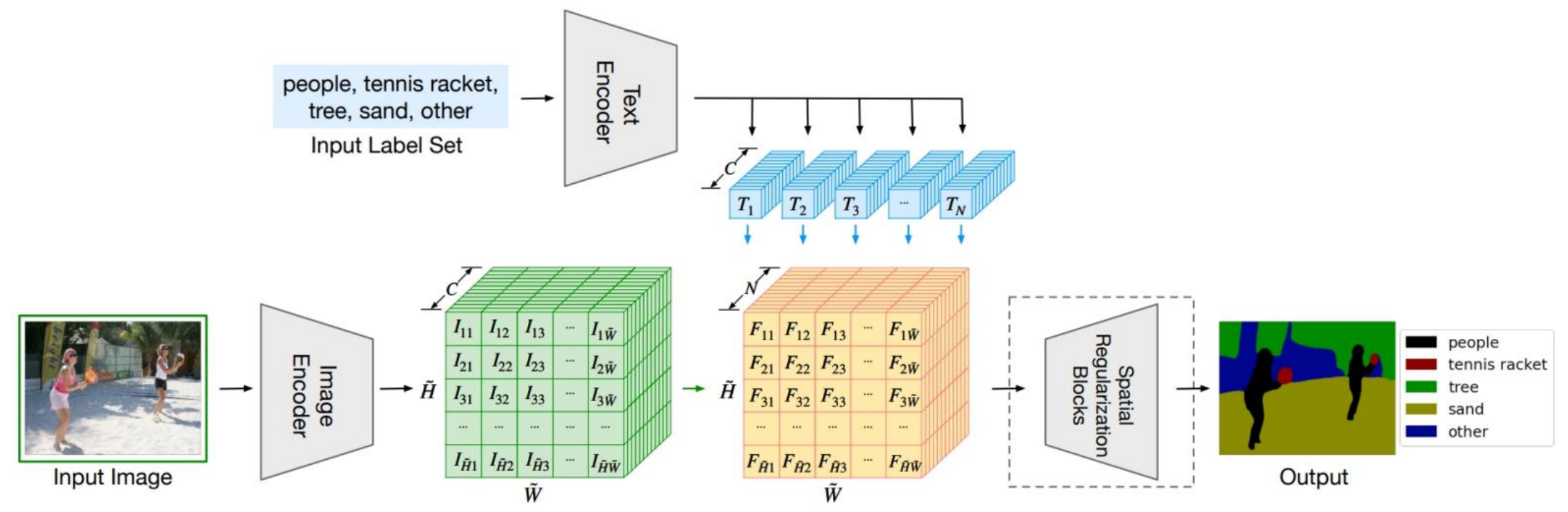
IV. Parameter-Efficient Fine-Tuning(PEFT)

PEFT techniques aim to reduce the number of parameters that need to be updated during transfer learning, thereby increasing the efficiency and effectiveness of model finetuning. When fine-tuning large pre-trained models, adjusting only a small number of parameters can achieve good performance, avoiding the high computational costs and memory requirements associated with full model finetuning.

Methods

➢ Pixel-Level Feature Matching

LSeg:



➢ Two-Stage Region-Level Image-Text Matching

OpenSeg

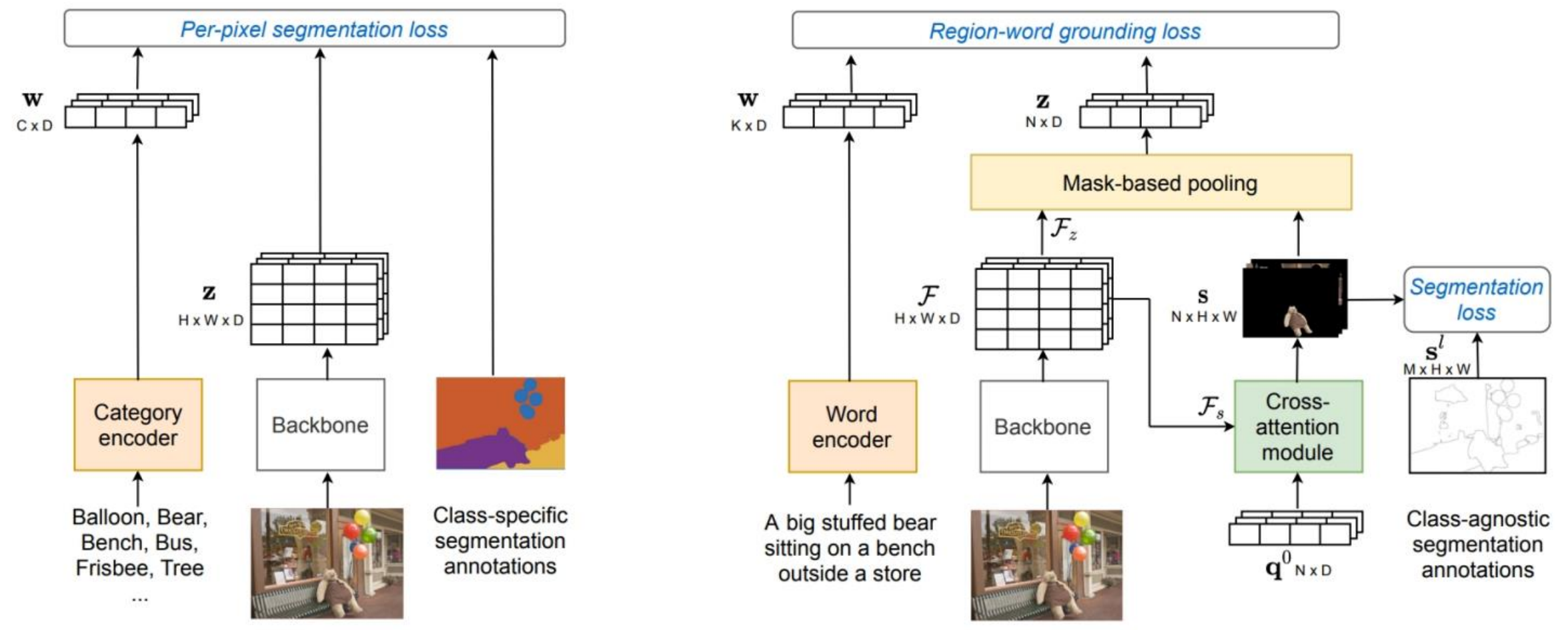

Per-pixel segmentation loss
W
C x D
Z
H x W x D
Category encoder
Backbone
Balloon, Bear, Bench, Bus, Frisbee, Tree ...
Class-specific segmentation annotations
Region-word grounding loss
W
K x D
Z
N x D
Mask-based pooling
$\mathcal{F}_z$
$\mathcal{F}$
H x W x D
S
N x H x W
Segmentation loss
S^l
M x H x W
Word encoder
Backbone
$\mathcal{F}_s$
Cross-attention module
A big stuffed bear sitting on a bench outside a store
q^0 N x D
Class-agnostic segmentation annotations

Figure 2. The difference between region-level feature matching and pixel-level feature matching.

OVSeg

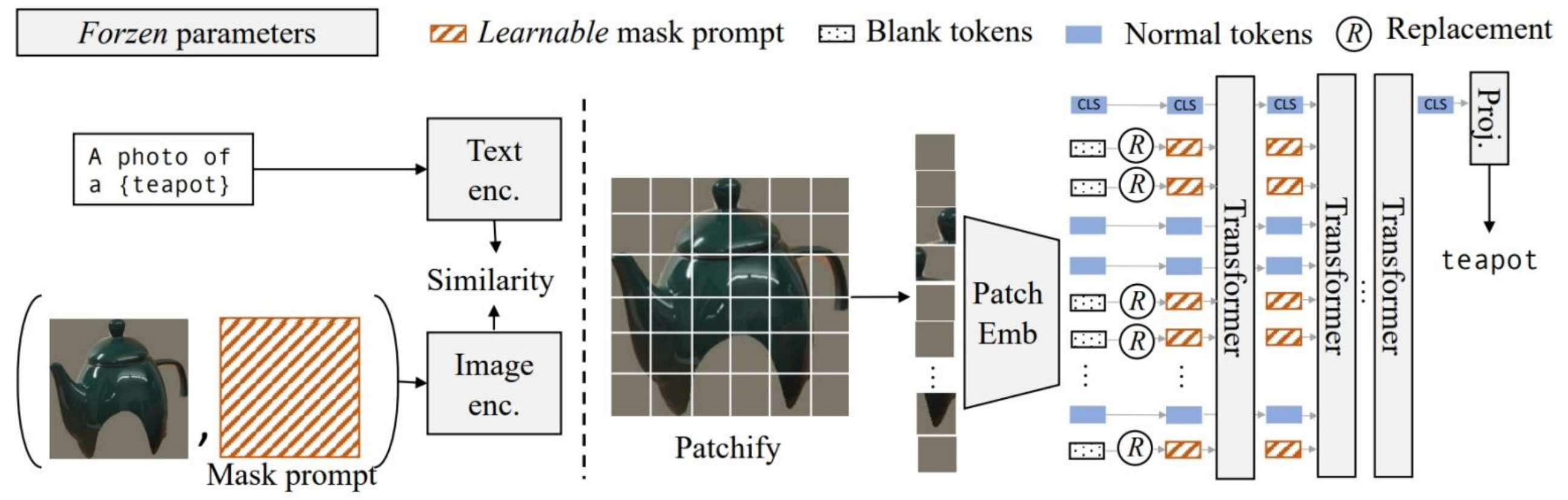


Figure 3. The Overview of Mask Prompt Tuning.

➢ Improved Single-Stage

SAN

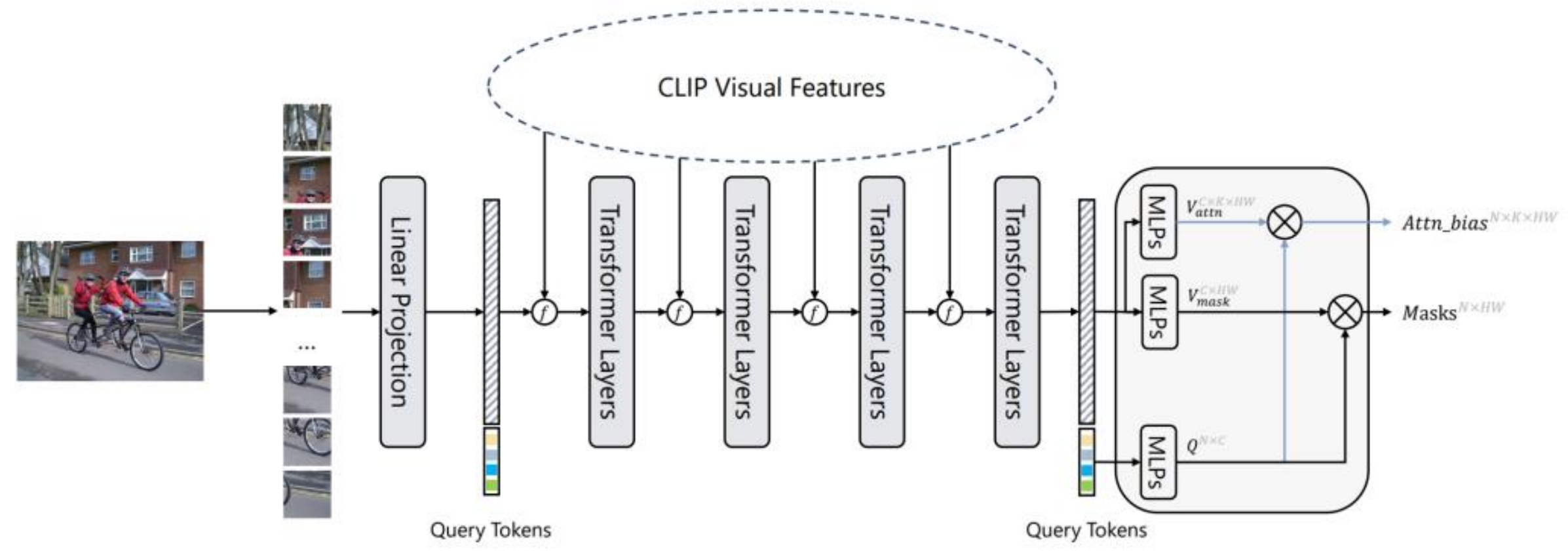


Figure 4. Overview of the SAN framework.

Challenges And Outlook

➢ Challenges

1. **Base class overfitting problem**

   Most methods rely heavily on base class annotation data to detect and segment objects. However, there is a natural gap in distribution and semantic information between new classes and base classes, affecting the accuracy of segmenting new class objects.

2. **Better benchmarks and metrics**

   The current datasets are still relatively small and insufficient for OVSS-related tasks.

3. **Training cost**

   Most state-of-the-art methods require a large amount of data for pretraining to achieve good performance.

➢ Future Directions

1. **Real-time open vocabulary detection and segmentation.**
2. **Combining large language models.**
3. **Unifying open vocabulary tasks.**

Conclusion

➢ In this survey, we have thoroughly examined the field of OVSS. In the background section, we briefly introduced the problem definition, history, preliminaries, datasets, and metrics of OVSS. In the methods section, we categorized the models into four types based on their architecture: Pixel-Level Feature Matching, Two-Stage Region-Level Image-Text Matching, Improved Single-Stage, and Open Segmentation. We then progressively introduced representative models from these four categories according to their technical routes and logical hierarchies, highlighting the research motivations and key technical points. Finally, we summarized the existing challenges and suggested future research directions for open vocabulary learning.