Unleashing the Potential of the Diffusion Model in Few-shot Semantic Segmentation

Muzhi Zhu[1], Yang Liu[1], Zekai Luo[1], Chenchen Jing[1], Hao Chen[1*], Guangkai Xu[1], Xinlong Wang[2], Chunhua Shen[1,3]

[1]Zhejiang University [2]Beijing Academy of Artificial Intelligence [3]Ant Group

Motivation

Our work reconsider the most fundamental question in using generative models for visual perception: **how to design a fine-tuning framework that can guarantee both generalization ability and precise prediction of details?**

Few-shot Semantic Segmentation (FSS) aims to segment query images given support samples. The demands of this task for open-set generalization and high-quality segmentation results precisely align with this challenge. **Thus, our motivation is to further address the fundamental question posed above by exploring the Diffusion Model on the FSS task.**

Method

Our investigation into model design primarily adheres to **two criteria**:

1. Simple and efficient as possible.
2. Maximize the preservation of the Latent Diffusion Model's generative schema.

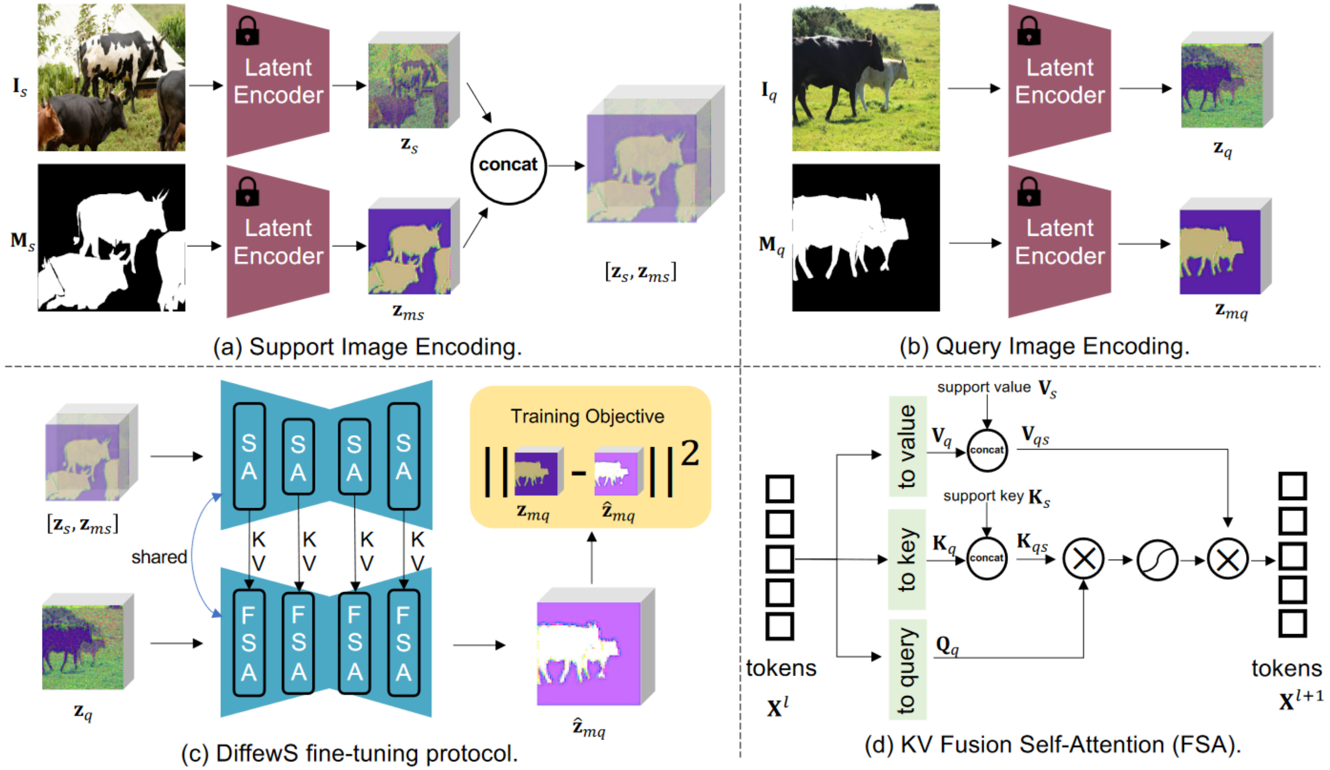
Specifically, **four key issues** need to be addressed:

1. How to facilitate interaction between the query image and support image?
2. How to effectively incorporate information from the support mask?
3. What form of supervision from the query mask would be most reasonable?
4. How to design an effective generation process to transfer the pre-trained diffusion models to mask prediction task?

Method

- **Interaction between query and support images**

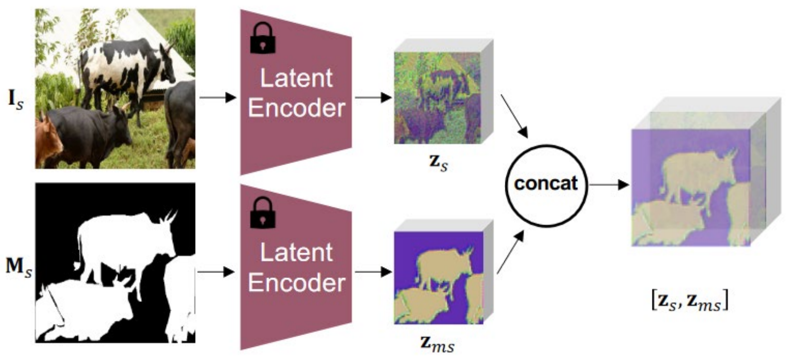
We first propose a **KV fusion method in self-attention layer (FSA)** to achieve interaction between query image and support image. Since we only replaced K and V, we can fully reuse the weights of the original self-attention.



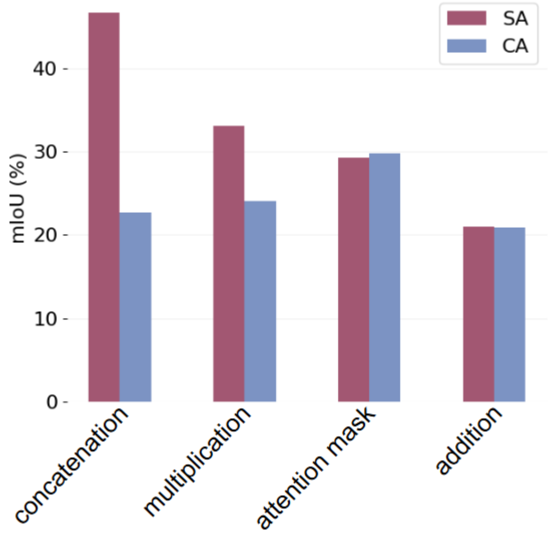
Method

- **Injection of support mask information**

Building upon the Self-attention KV fusion approach, we investigate four methodologies for incorporating support mask information. We observe that **Concatenation method** is surpassed the other three.



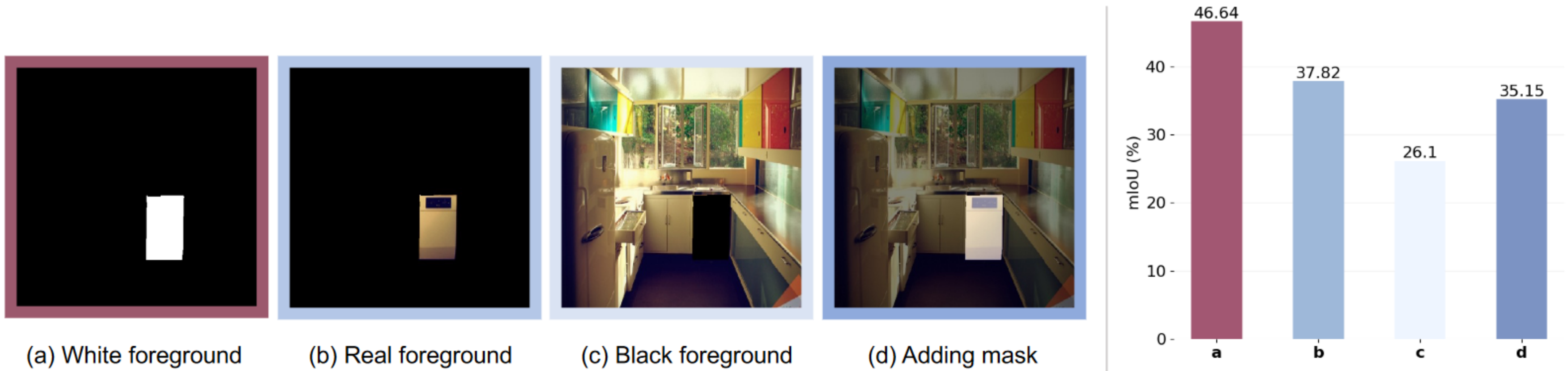
(a) Support Image Encoding.



Method

- ## Supervision from query mask

We explore four forms of supervision methods that balance ease of learning for the UNet and convenient post-processing for segmentation results. We find that directly using white **foreground + black background** achieves the best performance in all experiments.



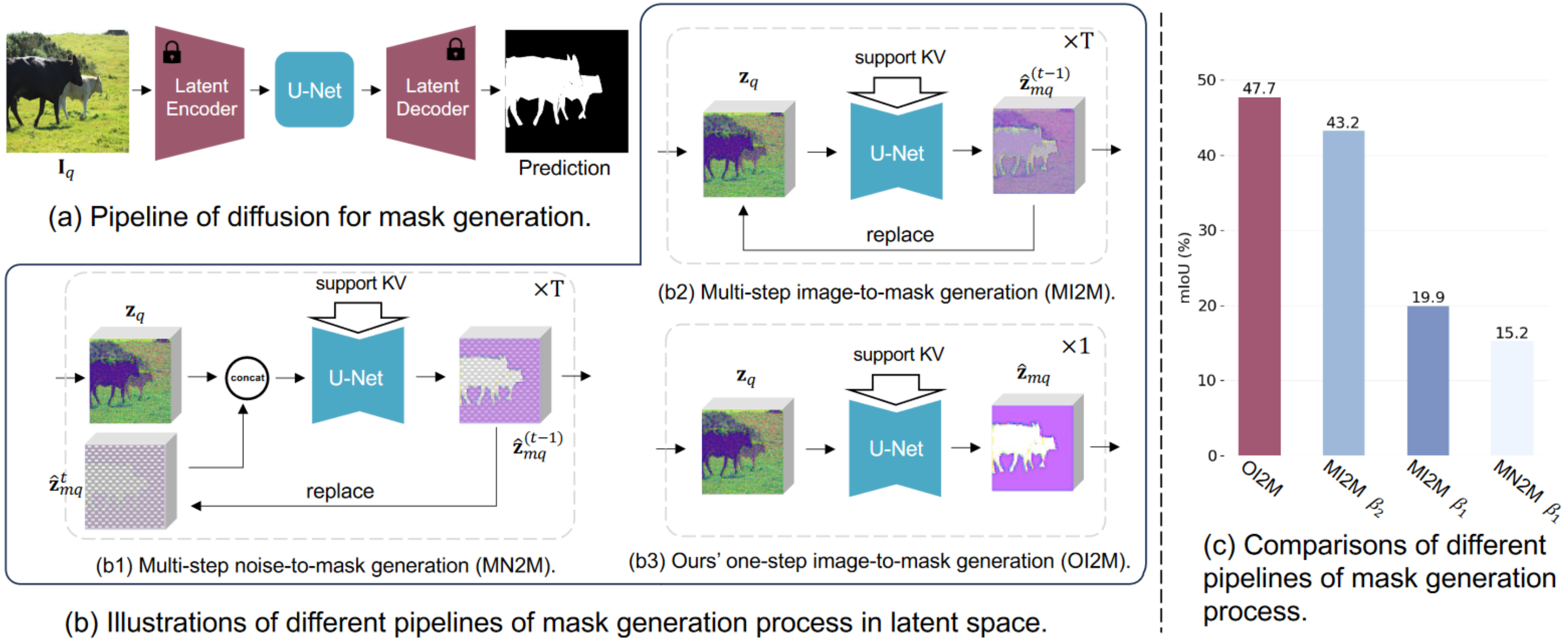
Comparison of different supervision.

Method

ZHEJIANG UNIVERSITY 1897 | BAAI 智源研究院 | 蚂蚁集团 ANT GROUP | NEURAL INFORMATION PROCESSING SYSTEMS

• Exploration of generation process

We explore three different mask generation processes and find that OI2M achieves best performance and improves the predictive efficiency.



(a) Pipeline of diffusion for mask generation.

(b1) Multi-step noise-to-mask generation (MN2M).

(b2) Multi-step image-to-mask generation (MI2M).

(b3) Ours' one-step image-to-mask generation (OI2M).

(b) Illustrations of different pipelines of mask generation process in latent space.

(c) Comparisons of different pipelines of mask generation process.

Experiment

- **Quantitative results**

Table 1 – Results of few-shot semantic segmentation on COCO-20^i, PASCAL-5^i, and LVIS-92^i, under in-context setting.

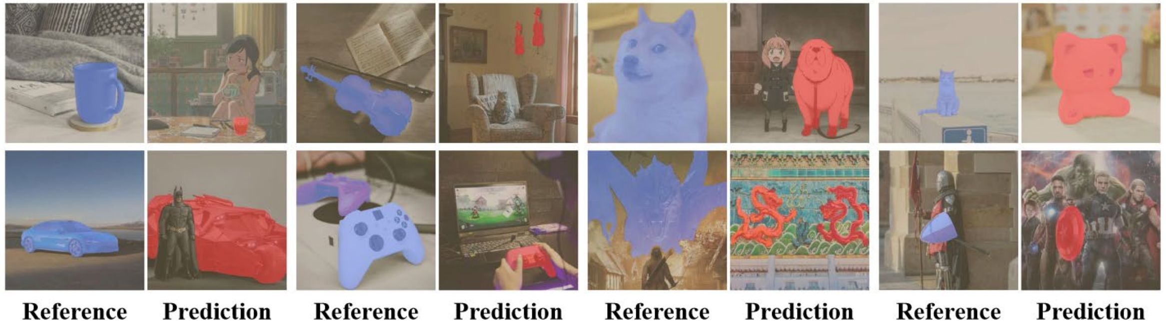
Methods	Venue	COCO-20^i one-shot	COCO-20^i few-shot	PASCAL-5^i one-shot	PASCAL-5^i few-shot	LVIS-92^i one-shot	LVIS-92^i few-shot
HSNet [18]	ICCV'21	41.7	50.7	68.7	73.8	17.4	22.9
VAT [47]	ECCV'22	42.9	49.4	72.4	76.3	18.5	22.7
FPTrans [48]	NeurIPS'22	56.5	65.5	77.7	83.2	-	-
Painter [29]	CVPR'23	32.8	32.6	64.5	64.6	10.5	10.9
SegGPT [30]	ICCV'23	56.1	67.9	83.2	89.8	18.6	25.4
PerSAM [49]	ICLR'24	23.0	-	-	-	15.6	-
PerSAM-F [49]	ICLR'24	23.5	-	-	-	18.4	-
Matcher [22]	ICLR'24	52.7	60.7	67.9	75.6	33.0	40.0
DiffewS	this work	71.3	72.2	88.3	87.8	31.4	35.4

Table 2 – Results of strict few-shot semantic segmentation on COCO-20^i.

Methods	Venue	1-shot 20^0	1-shot 20^1	1-shot 20^2	1-shot 20^3	1-shot mean	5-shot 20^0	5-shot 20^1	5-shot 20^2	5-shot 20^3	5-shot mean
HSNet [18]	ICCV'21	37.2	44.1	42.4	41.3	41.2	45.9	53.0	51.8	47.1	49.5
CyCTR [50]	NeurIPS'21	38.9	43.0	39.6	39.8	40.3	41.1	48.9	45.2	47.0	45.6
VAT [47]	ECCV'22	39.0	43.8	42.6	39.7	41.3	44.1	51.1	50.2	46.1	47.9
BAM [51]	CVPR'22	43.4	50.6	47.5	43.4	46.2	49.3	54.2	51.6	49.6	51.2
DCAMA [19]	ECCV'22	49.5	52.7	52.8	48.7	50.9	55.4	60.3	59.9	57.5	58.3
HDMNet [20]	CVPR'23	43.8	55.3	51.6	49.4	50.0	50.6	61.6	55.7	56.0	56.0
DiffewS	this work	47.7	56.4	51.9	48.7	51.2	52.0	63.0	54.5	54.3	56.0

Experiment

- **Qualitative results**



Thank You