AAAI-24 / IAAI-24 / EAAI-24
Edge Intelligence Workshop 2024

Self-Supervised Vision Transformers Are Efficient Segmentation Learners for Imperfect Labels

Seungho Lee*[1], Seoungyoon Kang*[1,2], Hyunjung Shim†[2]

[1]*Yonsei University*, [2]*Korea Advanced Institute of Science & Technology*

* indicates an equal contribution
† indicates a corresponding author

Introduction

Semantic Segmentation

- Essential for understanding each pixel in an image, widely used in autonomous driving and medical imaging.
- High precision required, making data preparation time-consuming and costly.

Approach

- Utilizing self-supervised vision transformers (SSVT) to effectively work with imperfect labels (scribble, point-level and image-level).
- Lowering annotation costs significantly while maintaining high accuracy.

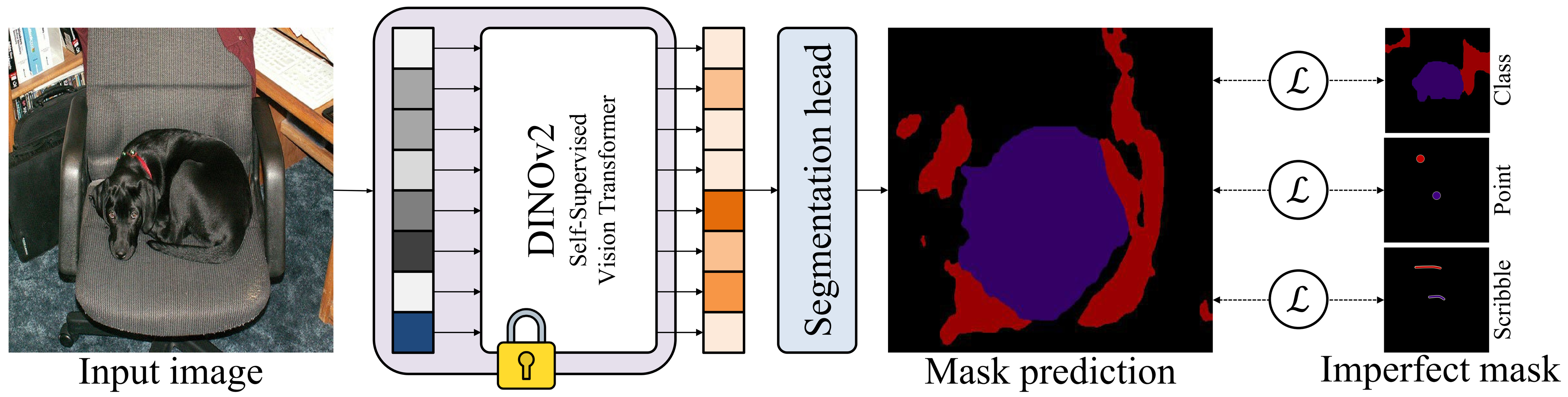
Research focus

- (1) Preserving structural information in SSVT for better segmentation results.
- (2) Only training the lightweight segmentation head, reducing overall training costs.

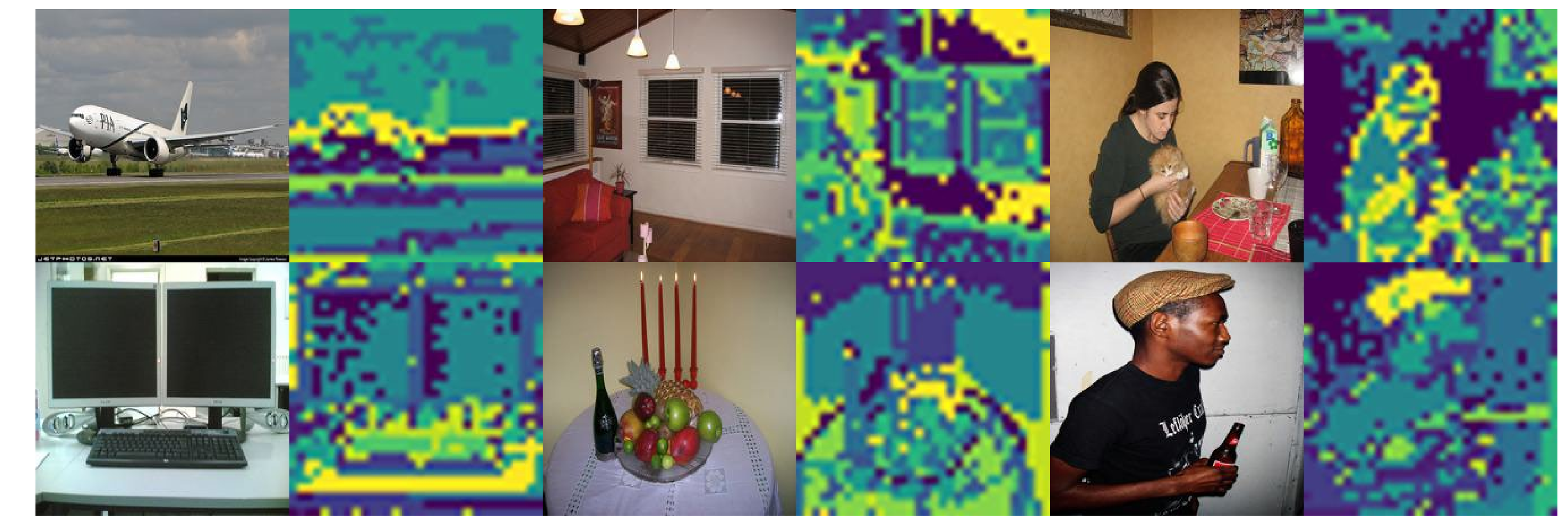
Contributions and results

- A cost-effective, generalizable approach for semantic segmentation under various imperfect label conditions.
- Outperformed existing methods, especially effective with text-driven labels from VL models (11.5%p).

Method



Overview of our method. $\mathcal{L}$ represents the matching loss for each imperfect mask type. For image-level label (class), $\mathcal{L}$ is pixel-wise cross-entropy. For others, $\mathcal{L}$ is masked pixel-wise cross-entropy. The backbone of the self-supervised vision transformer model is fixed during semantic segmentation training. Only the segmentation head is trained on imperfect masks and their corresponding images.



DINOv2 feature analysis. For each image pair, the right image is the result of applying K-means clustering to each token from DINOv2 using the left image. Without any supervision, DINOv2 exhibits a strong shape prior, indicating that the objects are identifiable only with the K-means clustering.

Experiments

	Baseline		Ours
Scribble	TEL'22	77.6	80.1
Point	TEL'22	68	73.6
Class (Image-level)	ADELE'22	69.3	71.2
	SegFormer'21	65.6	
Zero-shot VL	SegFormer'21	26.9	38.4

Quantitative evaluation of different types of imperfect label type. The cost of labeling decreases in the following order for each type of supervision: scribble, point, class (image-level), and zero-shot VL.

Method	GT	SEAM	EPS
Pseudo-label	-	63.6	69.4
DeepLabV1	75.8	64.5	70.1
DeepLabV3+	78.5	63.3	68.6
SegFormer	82.8	65.5	69.0
Ours	80.6	71.2	74.1

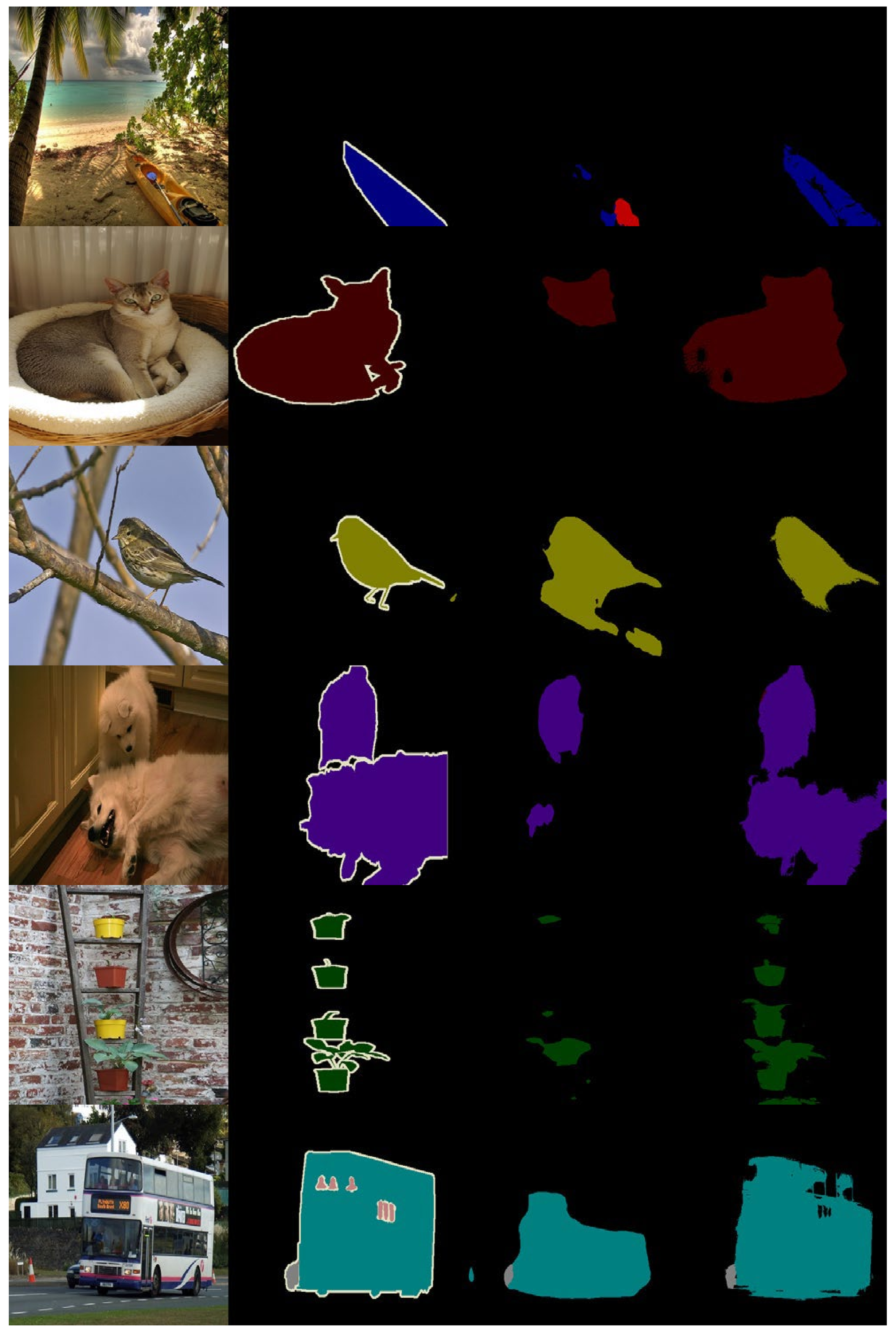
Image-level label quality-based performance comparison. Quality indicates the mIoU between the pseudo-label of each method and the ground-truth. For each method, we evaluate mIoU along various types of pseudo-labels used for training the segmentation model.

Method	Pretraining	Backbone strategy	
		Freezing	Tuning
DeepLabV1	Classification	64.6	64.5
DeepLabV3+	Classification	61.7	63.3
SegFormer	Classification	63.6	65.2
DINOv2 (ours)	Self-supervised	71.2	64.5

Performance analysis on backbone training strategies. Classification indicates model pretraining using ImageNet dataset.

Method	SEAM	EPS
Pseudo-label	63.6	69.4
DINOv1	58.9	63.6
ibot-L	65.2	70.0
ibot-L/22k	65.8	73.3
DINOv2	71.2	74.1

Self-supervised vision transformer performance across varying levels of imperfect label quality. All SSVT models are trained using our same strategy.



Qualitative evaluation on image-level labels.