
Thanks for your careful and valuable comments. We will explain your concerns point by point.

1 REVIEWER01

Q1: Abut the English grammar mistakes in the paper.

A1: Thanks for your correction. We have carefully proofread the article and have made the corresponding revision and make sure the use of symbols is consistent.

Q2: No ablation experiments for Lsc and Lr, which are described as sub-sections of the technical contribution in the approach.

A2: We have add the ablation experiments for Lsc and Lr for Deeplabv3+ and SwinZS3 in table1. Thanks for your advice.

2 REVIEWER02

Q1: The text flow feels irregular.

A1: We have carefully proofread the article and have made the corresponding revision and make sure the use of symbols is consistent.

Q2: Unseen-10 split results seem to be missing for both Pascal VOC and Pascal Context datasets.

A2: We put the results below. Due to papas limitations, we did not put the results in the camera-ready paper. Though the Unseen-2 -8 split results are enough to illustrate its performance, we consider adjusting the paper for adding the Unseen-10 split results. Thanks for your advice.

Table 1: Quantitative results on the PASCAL VOC and Context validation sets.

K	method	VOC			Context		
		$mIoU_s$	$mIoU_u$	$hIoU$	$mIoU_s$	$mIoU_u$	$hIoU$
10	DeViSE	31.7	1.9	3.6	17.5	1.3	2.4
	SPNet	59.0	18.1	27.7	27.1	9.8	14.4
	ZS3Net	33.9	18.1	23.6	20.8	12.7	15.8
	CSRL	59.2	21.0	31.0	29.4	14.6	19.5
	JoEm	63.5	22.5	33.2	33.0	14.9	20.5
	Ours	64.7	23.8	34.2	33.8	15.9	22.2

Q3: There is no clear experimental result to support the central claim that the seen/unseen bias is alleviated via minimizing the euclidean distance and using the pixel-text score maps.

A3: Actually, whether the seen/unseen bias problem is partially solved should look at the change of SwinZS3 mIoU/hIoU. And we give a explanation in Figure.3. The pixel-text score maps and its decision boundary show better performance than not using it.

3 REVIEWER03

Q1: The structure of Language Encoder.

A1: The Language Encoder which we use in paper is just an MLP layer. We have add this explanation on our paper. Thanks for your advice!

Q2: There are many problems with the layout and writing of the article. For example, there is incorrect capitalization in the first line of section 3.1 and missing spaces in line 4. There is confusion about the case of symbols in section 3.4.

A2: We have carefully proofread the article and have made the corresponding revision and make sure the use of symbols is consistent.

Q3: Some recent related papers are not cited. Such as [1, 2, 3]

A3: These papers are cited in the new version. Thanks!

4 REVIEWER04

Q1: It seems that the paper claims its novelty as using Swin Transformer, however, Swin Transformer has been widely used in CV community, and simply adopting it as the backbone is not novel.

A1: Actually, it is not easy to use the swin transformer to ZS3. There are at least three questions to get good reasonable and good results.

(1) For the pixel-text feature aligning work, what the network should be modified? For this, we did modify the swin-transformer pooling layer from avg pooling to max pooling. Because the avg pooling will cause the semantic shift problem, and we give a visual results for explaining this problem. For the zero-shot pixel-text framework, the avgpooling will cause the foreground feature and background feature shift. And we think the maxpooling could alleviate this problem.



Figure 1: left :maxpooling right: avgpooling

(2) How to pretrain the swin-transformer? There are at least three chooses. 1. Using the common Imagenet pretrained weight. It is not good for the supervision leakage. 2 Using the Imagnet dataset and removing the unseen classes labeled images. 3 Using the self-supervised weight. We did a lot of experiment for choosing the self-supervised weight.

(3) The super-parameters setting.

So, it is a basic but necessary work for adopting the swin-transformer to the zero-shot semantic segmentation. And the cross-entropy loss, regression loss, semantic consistency loss is not the key points in our paper, we never claim that this is our innovation. Though the pixel-score map is used

in some works, the pixel-score decision boundary is never used in previous zero-shot works. We hope you could consider this job and your score more seriously. Thanks for you very much.

Q2: The comparisons of the experiments (Table. 2) might be unfair. It uses stronger backbone (Swin Transformer) network compared to other methods. Please add experiments of DeepLabV3+ for all K in Table. 2.

A2: In table.2 all the others use the Deeplabv3+ actually. The JoEm is one of the typical approach for discriminative zero-shot semantic segmentation models using Deeplabv3+. And for the approach we proposed, the ablation table.1 gives a reasonable compare between Deeplabv3+ and swin-transformer.

Q3: Please also discuss the relationship with CLIP-based zero-shot segmentation methods in the related work section.

A3: We mentioned the CLIP-based methods in our introduction. Actually, we argue that the CLIP-based methods are not really zero-shot. Because we have used the labeled images or object for training the CLIP models. But the zero-shot model should never see the image-class label. So, for word2vec ZSSS works, it is not necessary to cite the CLIP-based methods in related work.

Q4: The paper writing needs significant improvement and careful revision. (1) The use of symbols is inconsistent. Does s in Eq.4 and Eq.6 the same? Some times N is used for the number of seen categories, sometimes K is used. (2) The introduction of "regression loss" is mostly unclear. where does semantic feature maps (s) come from? (3) The use of citation is wrong throughout the whole paper. (4) There are many typos. for example, extra "(" in Table. 2; "deeplabv3+" and "Deeplabv3+" in Table. 1.

A4: We have carefully proofread the article and have made the corresponding revision and make sure the use of symbols is consistent.