Notes on revision made to manuscript Paper #TCDS-2023-0486

We thank the editors and reviewers for their constructive comments and valuable suggestions. The manuscript has been thoroughly revised according to the editor and reviewers' comments. As suggested by the Editor-in-Chief, the previous version of manuscript #TCDS-2023-0486 is revised and resubmitted as a new submission. The comment-wise responses are given below:

Response to the Comments of Associate Editor

Associate Editor Comment — This paper presents a novel method called DatUS2 for unsupervised semantic segmentation (USS). The main idea is to leverage patch embeddings derived from a pre-trained vision transformer (ViT) to break down a scene into multiple segments. Overall, this idea is novel, compared with previous studies. However, both reviewers raised some valid concerns, including the feasibility of the pseudo labeling, some experimental set-up, and details, as well as some ablation studies. The authors are suggested to address all the concerns raised by the reviewers in the revision.

Response: We have addressed all the reviewer's concerns, including the feasibility of the pseudo labeling, experimental setup, details, and ablation studies. Also, we have made substantial changes in our revised manuscript to accommodate reasonable comments from the reviewers.

Response to the reviewers

Reviewer 1

This manuscript introduces $DatUs^2$, an unsupervised semantic segmentation (USS) as the downstream task of a pre-trained self-supervised vision Transformer (ViT). Experiments show that (1) The USS task could be recognized as an evaluation metric of self-supervised methods; (2) The proposed $DatUs^2$ outperforms existing methods by a large margin in the USS task. However, we have some serious concerns with the paper as follows and suggest authors to solve them before resubmitting:

Reviewer Comment 1.1 — I can not understand whether the manuscript aims to propose a novel USS method or introduce the USS task as an evaluation of self-supervised ViT. At the beginning of the abstract, the authors mentioned that the USS is a good evaluation for self-training schemes but later the author said that $DatUS^2$ is a novel USS method and lists experimental results. What is the main goal of this article? Whether the manuscript introduces the USS as an evaluation of self-supervised methods or proposes a novel USS method?

Response: The main objective of the paper is to introduce a novel downstream task $DatUS_2$ to evaluate existing self-supervised training schemes for the unsupervised dense semantic segmentation task. The experiment section (section IV of the revised manuscript) provides a detailed comparison of

the performance of existing self-supervised training schemes for the proposed downstream task. Also, the proposed downstream task $DatUS^2$, along with a self-supervised vision transformer, is a novel Unsupervised Semantic Segmentation (USS) method. Hence, we also compare it with the existing state-of-the-art USS methods.

The abstract, introduction, and conclusion sections are revised to state the main objective of the manuscript clearly.

Reviewer Comment 1.2 — The 'Semantic-wise Pseudo Labeling' method is unreasonable. For example, if there are multiple objects of multiple categories on an image, and these objects do not overlap with each other, it is impossible to segment these objects with the k-means algorithm, where K is a fixed number and equal to the category number in the manuscript.

Response: We understand the reviewer's concern, but our 'Segment-wise Pseudo Labeling' step differs from directly applying k-means clustering on the single image's pixels. The 'Segment-wise Pseudo Labeling' step is proposed to overcome the exact issue raised by the reviewer about the pixel clustering of a single image with the k-means algorithm. It is important to note that the 'Segment-wise Pseudo Labeling' step is performed globally (on the entire dataset) in two stages. It applies the k-means algorithm (in the second stage) to the segments obtained from all images in the previous steps. Next, we explain "how the 'Segment-wise Pseudo Labeling' step overcomes these issues?':

As the reviewer correctly pointed out, it is impossible to directly segment a single image into the non-overlapping objects of multiple categories with k-means clustering. Hence, we do not apply k-means directly on the pixels of a single image. We overcome such limitations with the following steps:

- We first decompose each image from the dataset into various segments in the first three steps of the proposed method (refer to section III(a),(b), and (C) of the revised manuscript). In this step, we utilize a self-supervised vision transformer and an unsupervised graph clustering algorithm (Louvain Clustering [1]).
- Next, we apply two-stage self-supervised image classification to cluster all image segments from the dataset (refer to section III(D) of revised manuscript). In the first stage, the visual feature of each segment is extracted from the self-supervised backbone. Figure 9 of the revised manuscript shows that the segments of the same visual group are highly correlated in the visual feature space. In the next stage, the k-means clustering algorithm is trained with visual features set to assign cluster id $[0, K]$ to corresponding segments, where $length(visualfeatureset) >> K$.
- Finally, in the fourth step (refer to section III(E) of revised manuscript), the pseudo mask is generated by assigning the cluster-ID $'k'$ of the segment to the image pixel within the segment (refer to section III(d)).

Also, the number of segments is at least three times the size of the dataset, which is much bigger than the fixed number of clusters, i.e., K in k-means clustering. So, the 'Segment-wise Pseudo Labeling' step does not suffer from the issue raised by the reviewer.

Reviewer Comment 1.3 — The mathematical symbols and capitalization are very messy and frustrating to read. [1]Please avoid using consecutive capital words or phrases in the middle of a sentence; if necessary, use initialisms. [2] Please carefully check all the mathematical symbols in the manuscript. It is really messy. For example, in P5 L56 (refer to last line of the second paragraph

of section III(D) in the revised manuscript) and P6 L48 (refer to second line of the third paragraph of section III(D) in the revised manuscript), the Greek alphabet tau, i.e., τ has been defined twice. [3] in P7 L55 (refer to the first paragraph of section IV(D), the numbering in the list is incorrect.

Response: The manuscript has been revised as follows:

1. The consecutive capital words or phrases used in the manuscript have been replaced with corresponding abbreviations. Also, a dedicated table, which lists key abbreviations with expansion, is included in the revised manuscript (refer to Table 1 of the revised manuscript).

2. All the mathematical symbols and numbering are thoroughly checked and corrected accordingly. Also, a dedicated table, which lists key notations with their definition, is included in the revised manuscript (refer to Table 2 of the revised manuscript).

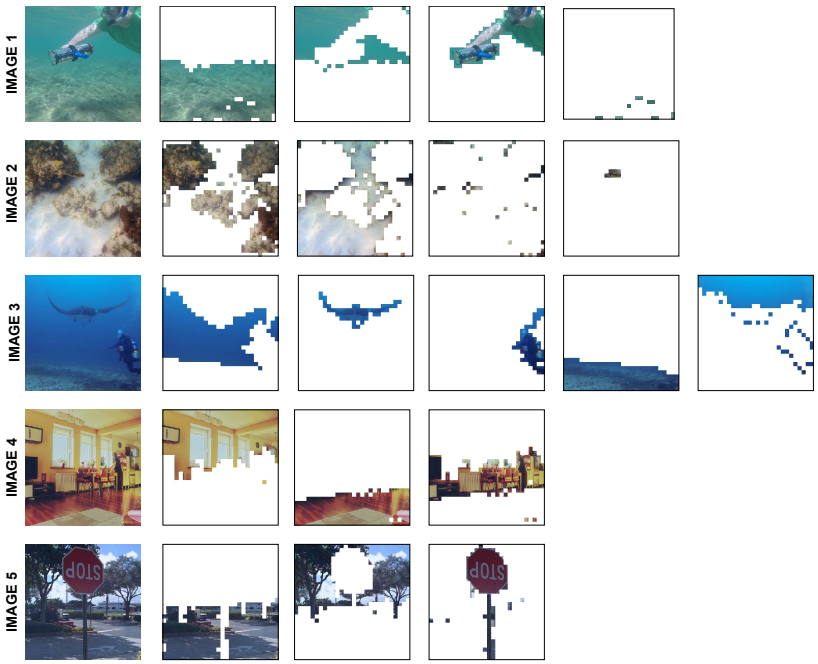

Figure 1: Visual demonstration of declaring valid and invalid segments of an image in the *Discover Image Segment* step (refer to section III(C) of the revised manuscript) of the proposed method.

Reviewer Comment 1.4 — In the experiments part, Table II (refer to Table IV of the revised manuscript), how to define the valid segment? A more clear illustration is appreciated.

Response: As per the suggestion, a clear definition of a valid segment is provided in the second paragraph of section III(D), and the second paragraph of section IV(D) of the revised manuscript.

Also, each valid segment of an image discovered in the third step of our proposed method captures an object from the scene. Also, as shown in Figure 9 of the revised manuscript, each valid segment is a meaningful collection of neighbor patches from an image. Figure 1 shows the segments of a few images discovered in the third step of the proposed work (refer to section III(c) of the revised manuscript) without further processing (to separate disconnected segments). It can be observed that **IMAGE 1** of

Figure 1 is majorly decomposed into three parts: river bed, water, and diver. Apart from that, a few disconnected segments are there, which are just a tiny portion of these three major segments. Such small segments are made of few patches and do not contribute to further steps. Hence, we declare major segments as valid and others as invalid segments. Similarly, **IMAGE 2** of Figure 1 is majorly decomposed into rock and river beds. Other tine segments can be declared invalid for further processing. Similar examples can be seen in the other images shown in Figure 1.

Also, as explained in section IV(B) of the revised manuscript, "Based on the qualitative observation, we consider the segments with more than five patches valid for pseudo labeling, i.e., the parameter $\tau = 5$, irrespective of the model and patch sizes."

Reviewer Comment 1.5 — An ablation study of the distribution of seed points of the k-means algorithm should be added. The distribution of seed points significantly influences the clustering result.

(a) SUIM Dataset (b) SUIM Dataset (c) SUIM Dataset

(d) COCO Dataset (e) COCO Dataset (f) COCO Dataset

Figure 2: The effect of the distribution of seed point (initial centroids) on clustering results (MIoU, Pixel Accuracy, and Avg. F1 Score). The sub-figures (a),(b), and (c) show ablation results on the SUIM dataset. Similarly, sub-figures (d), (e), and (f) show ablation results on the COCO dataset.

Response: In response to this comment, we perform an ablation study to analyze the effect of the distribution of seed points (initial centroids) of the k-means algorithm on clustering results. We choose three popular ways of initializing the centroids before the k-means clustering: (1) *k-means++ Initialization [2]* (utilizes the empirical probability distribution of points), (2) *Uniform Initialization* (uniformly sample the required number of centroids from the feature space), and (3) *Random Initialization* (randomly initialize the centroid with the same dimension as a point in the feature space).

The sub-figures in Figure 2 plot the clustering metrics (MIoU, Pixel Accuracy, and Avg. F1 Score for three initialization techniques (k-means++, Uniform, Random). The sub-figure (a), (b), and (c)

show results for the SUIM dataset. Similarly, plots (d), (e), and (f) show results on the COCO dataset. The results show that the k-means++ Initialization provides superior and consistent performance with increasing clusters (K) for both datasets. Also, uniform Initialization has competitive performance, but random Initialization is not recommended. Section IV(B), i.e., the Setup subsection of the revised manuscript is updated accordingly.

Reviewer 2

This paper introduces a new data-driven method, named $DatUS^2$, for unsupervised semantic segmentation, designed for use as a downstream task.

- The main approach involves leveraging patch embeddings derived from a pre-trained self-supervised vision transformer to break down a scene into multiple segments, a critical step for comprehensive scene segmentation, which is new compared with prior works.
- The paper has also demonstrated that the suggested unsupervised dense semantic segmentation method can serve as a downstream task for assessing self-supervised training schemes for vision transformers within a completely unsupervised context.
- In the experiments, a noticeable improvement was observed, compared with the benchmark method STEGO (ViT-S/8) on the public dataset SUIM.

I think the method is new and has contributions to the field. My comments are mainly about the results and experiments:

Reviewer Comment 2.1 — In Table V (refer to Table VII of the revised manuscript), the proposed paper is compared with other baselines using the ViT-B or ViT-C as the backbone. The results show some improvement under such a setting. The paper could be more solid if showing results using other backbones (e.g., Deit, Swin, etc.), as ViT-B and ViT-C are already not the SOTA transformer architecture.

Response: This primary objective of our work is to propose a novel downstream task $DatUS^2$ to evaluate the existing SOTA self-supervised training schemes [3–7]. The proposed $DatUS^2$ method works upon a vision transformer already pre-trained with a self-supervised training scheme. Hence, we only consider the vision transformer architectures compatible with these SOTA self-supervised training schemes, and their pre-trained weights are publicly available. Self-supervised pre-training and optimization of other existing supervised vision transformer architectures are outside the scope of our research work. Thus, they may not be a fair competitive backbone for the proposed work.

Also, ViT-based architecture like Deit and Swim are supervised models and propose architectural, training, or data-based optimization to achieve superior results. In our research work, we consider the DINO [3], DINOv2 [4], and other SOTA self-supervised training schemes [5–7], which already utilize the advancement proposed by novel ViT-based methods like Deit, and Swim [8, 9] for training. The self-supervised training scheme of the DINO framework utilizes the architectural and training methodology of the Deit model to improve the performance of the ViT model in an unsupervised setting. For example, DINO's student-teacher training strategy is motivated by the Deit framework. Also, the ViT architecture of DINO follows the implementation of the Deit model.

In summary, ViT-based architectures like Deit and Swin are optimized for task-specific supervised training and may not be compatible with the proposed method. We may not directly utilize such ViT backbones in unsupervised experiments. Also, in the future, the proposed method can be used to evaluate novel self-supervised training schemes that directly utilize Deit or Swim models.

Reviewer Comment 2.2 — When compared with IIC and PiCIE, the proposed method used a different backbone (refer to Table VII of the revised manuscript). IIC used ResNet/VGG as the backbone. PiCIE used FPN+ResNet as the backbone. However, the backbone of the proposed method is ViT-B. Such a comparison is not fair and does not make too much sense. The authors should use the same backbone to do the comparison.

Response: The IIC [10] and PiCIE [11] are some of the pioneering works proposed for the dedicated unsupervised semantic segmentation task. These methods utilize a CNN backbone for self-supervised training and utilize pixel clustering for training convergence.

In addition to our primary purpose, we also compare our method with the existing unsupervised semantic segmentation tasks like PiCIE, IIC, and STEGO [12], unlike the PiCIE and IIC, we utilize a pre-trained ViT backbone [3–7], which enables us to perform graph clustering to discover valid segments of images form dataset. Later, we perform k-means clustering of valid segments to generate pseudo masks. The comparison results validate that performing clustering at the segment level can achieve superior results.

Reviewer Comment 2.3 — All the ablation studies contain the results of MIoU, PAcc, and Avg F1. However, the Avg F1 scores of the baseline methods are missing (refer to Table VII of the revised manuscript). Can the proposed method still outperform the baseline methods on the Avg F1 scores?

Response: We consider the benchmark metrics from existing work for comparison; they do not list their performance for the Avg F1 score. Irrespective of this, our method does outperform the baseline for the Avg. F1 score, and achieves 16.06% improvement on the SUIM dataset. Table VII of the revised manuscript has been updated with available results.

Reviewer Comment 2.4 — The ablation study will be more solid if the ablation results on the COCO dataset are provided.

Response: As per the suggestion, we have included the ablation results on the validation split of the COCO dataset in section IV(G) of the revised manuscript.

References

[1] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[2] D. Arthur, S. Vassilvitskii *et al.*, "k-means++: The advantages of careful seeding," in *Soda*, vol. 7, 2007, pp. 1027–1035.

[3] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[4] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] P. Zhou, Y. Zhou, C. Si, W. Yu, T. K. Ng, and S. Yan, "Mugs: A multi-granular self-supervised learning framework," *arXiv preprint arXiv:2203.14415*, 2022.

[7] X. An, J. Deng, K. Yang, J. Li, Z. Feng, J. Guo, J. Yang, and T. Liu, "Unicom: Universal and compact representation learning for image retrieval," *arXiv preprint arXiv:2304.05884*, 2023.

[8] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers amp; distillation through attention," in *International Conference on Machine Learning*, vol. 139, July 2021, pp. 10 347–10 357.

[9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[10] X. Ji, J. F. Henriques, and A. Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9865–9874.

[11] J. H. Cho, U. Mall, K. Bala, and B. Hariharan, "Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 794–16 804.

[12] M. Hamilton, Z. Zhang, B. Hariharan, N. Snavely, and W. T. Freeman, "Unsupervised semantic segmentation by distilling feature correspondences," *arXiv preprint arXiv:2203.08414*, 2022.