Supplementary Material: Revisiting the Encoding of Satellite Image Time Series

Xin Cai
cai-x@ulster.ac.uk
Yaxin Bi
y.bi@ulster.ac.uk
Peter Nicholl
p.nicholl@ulster.ac.uk
Roy Sterritt
r.sterritt@ulster.ac.uk

School of Computing
Ulster University
Belfast, UK

1 Network Instantiation

Because of the flexibility of SITS reformulation and the versatility of the proposed collect-update-distribute learning procedure, we chose to draw on recent advances in CV where object queries in the transformer decoder have been reinterpreted as cluster centres and cross-attention has been recast as a clustering operation [2, 12, 16, 18], reviving the classical idea of framing image segmentation as a pixel grouping procedure rather than per-pixel classification. As clustering is essentially a quantization process where redundant information is gradually filtered out and therefore abstract concepts or high-level semantics may emerge, it has the potential for generic representation learning, not only limited to image segmentation tasks, as demonstrated by the recent pioneering work [8, 17]. As the main focus of this paper is to establish an effective representation learning framework for SITS, we decided to borrow the core building unit Group Propagation Block (GP Block) from GPViT [17] to instantiate the idea, leaving the architectural invention for future work. We simply incorporate the construction of GP Block for completeness as follows and refer readers to the original work [17] for specific details:

$$\boldsymbol{C}^v = \text{Concat}_h\left(\text{Softmax}\left(\frac{1}{\sqrt{2d}}\boldsymbol{C}^v\boldsymbol{W}_h^Q\left(\boldsymbol{V}\boldsymbol{W}_h^K\right)^T + \frac{1}{\sqrt{2d}}\boldsymbol{C}^p\boldsymbol{U}_h^Q\left(\boldsymbol{P}\boldsymbol{U}_h^K\right)^T\right)\boldsymbol{V}\boldsymbol{W}_h^V\right) \tag{1}$$

where $\boldsymbol{W}_h^{Q,K,V}$ and $\boldsymbol{U}_h^{Q,K}$ are projection matrices for content and position embeddings, respectively. Eq.(1) implements the collection process by using cross-attention where the affinity matrix is calculated through scaled dot-product and the softmax function is used for selecting the most relevant temporal elements.

$$\begin{aligned} \boldsymbol{C}^v &= \boldsymbol{C}^v + \mathrm{MLP}_1\left(\mathrm{LayerNorm}\left(\boldsymbol{C}^v\right)^T\right)^T \\ \boldsymbol{C}^v &= \boldsymbol{C}^v + \mathrm{MLP}_2\left(\mathrm{LayerNorm}\left(\boldsymbol{C}^v\right)\right) \end{aligned} \tag{2}$$

Eq.(2) implements the context cluster updating by using a MLPMixer [13] with one MLPs operated along the token dimension and another MLPs operated along the channel dimension.

$$\begin{aligned} \boldsymbol{Z} &= \mathrm{Concat}_h\left(\mathrm{Softmax}\left(\frac{1}{\sqrt{2d}}\boldsymbol{V}\tilde{\boldsymbol{W}}_h^Q\left(\boldsymbol{C}^v\tilde{\boldsymbol{W}}_h^K\right)^T + \frac{1}{\sqrt{2d}}\boldsymbol{P}\tilde{\boldsymbol{U}}_h^Q\left(\boldsymbol{C}^p\tilde{\boldsymbol{U}}_h^K\right)^T\right)\boldsymbol{C}^v\tilde{\boldsymbol{W}}_h^V\right) \\ \boldsymbol{Z}' &= \mathrm{Concat}\left(\boldsymbol{Z},\boldsymbol{V}\right)\tilde{\boldsymbol{W}}_{proj} \\ \boldsymbol{V}' &= \boldsymbol{Z}' + \mathrm{FFN}\left(\boldsymbol{Z}'\right) \end{aligned} \tag{3}$$

where $\tilde{\boldsymbol{W}}_h^{Q,K,V}$ and $\tilde{\boldsymbol{U}}_h^{Q,K}$ are a different set of projection matrices for content and position embeddings, respectively, $\tilde{\boldsymbol{W}}_{proj}$ is for linear projection of the concatenated features to the same dimension as the input, and FFN is a feed-forward neural network. Eq.(3) implements the distribution process by using input temporal elements as queries to gather information from updated context clusters, performing cross-attention in the reversed direction.

2 Implementation Details

2.1 Classification

We train and validate the classification model on PASTIS pixelset format dataset. Based on the observation from [11, 15] that an additional MLP projector is beneficial for reducing the transferability gap between unsupervised and supervised pre-training, we append the projector proposed in t-ReX [11] after the feature extractor Exchanger and use cosine softmax cross-entropy loss. We use AdamW[7] optimizer, a batch size of 128, a weight decay of 0.005, an initial learning rate of 0.0002, and a step learning rate scheduler which decays the learning rate at 0.7 and 0.9 fractions of the total number of training steps by a factor of 10 to train models for 50 epochs on 4 V100 GPUs. We randomly drop temporal observations by uniformly sampling from the interval between 0.2 and 0.4 as a data augmentation strategy to counter the adverse effect of cloud obstruction, which has also been adopted in training semantic & panoptic segmentation models.

2.2 Semantic & Panoptic Segmentation

We then use the pre-trained model to initialize Exchanger which serves as the temporal encoder in the semantic/panoptic segmentation pipeline, unless otherwise specified. For the Unet [10] used as the spatial encoder, we use the AdamW[7] optimizer, a batch size of 4, a weight decay of 0.005, an initial learning rate of 0.0002, and a poly decay learning rate scheduler to train models for 100 epochs on 4 V100 GPUs with Focal cross-entropy loss [6] for semantic segmentation and Parcels-as-Points (PaPs) prediction head and PaPs Loss[5] for panoptic segmentation. As it cannot fit a single input SITS sample with a spatial resolution of 128×128 and the temporal length of more than 30 into V100 GPU with 16G memory, we perform random crop with a crop size of 32×32 in training and test the model performance on full resolution on a A100 GPU. For concatenating the Exchanger with Mask2Former[2] framework, we mainly follow the settings in [2] only with the learning rate changed to 2×10^{-5}. And we train models for 100 epochs with a random crop size enlarged to 64×64, a batch size of 1 on 8 V100 GPUs. Please note when evaluating Exchanger+Mask2Former for panoptic segmentation we split the input into four 64×64 patches and stitch the prediction results together [1].

3 Convergence Analysis

We demonstrate the successful transfer of the pretrain-finetune paradigm from CV to SITS analysis, which is enabled by the reformulated SITS representation, shifting from spatiotemporal signals to sets of instances. It allows the backbone network to be pre-trained on efficient pixel-set format and then fine-tuned on standard spatiotemporal grids for downstream dense prediction tasks. Specifically, as shown in Fig. 1, pre-trained Exchanger as backbone network appended with a commonly-used segmentation model Unet with randomly initialized weights has led to faster convergence, more stable training and higher validation accuracy than completely training from scratch.

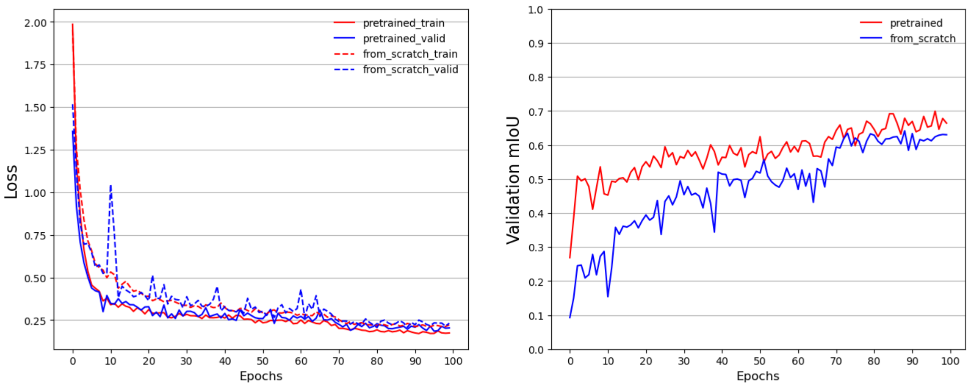

Figure 1: Convergence analysis for Exchanger+Unet with pre-trained backbones or training from scratch on PASTIS validation dataset (Fold-1). The left figure shows the training and validation losses. The right figure shows the evaluation metric mIoU on the validation dataset.

4 Color Palette for PASTIS

5 Visualisation of the Latent Features in Exchanger

We show latent features from the output of stage-1 and stage-2 of Exchanger before the projector head in Fig.3. It can be seen first that the intra-class variation is significantly reduced

[1] We found empirically that the panoptic evaluation metric is particularly sensitive to spatial resolution because of the spatial position encoding extrapolation and patch tokenization layer used in ViT [1, 3].

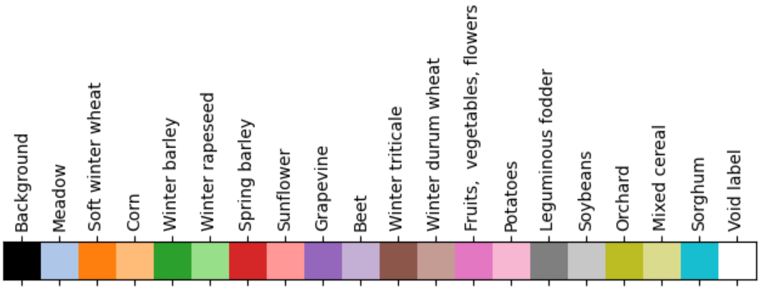

Figure 2: Color Palette used for visualising latent features, semantic & panoptic predictions on PASTIS.

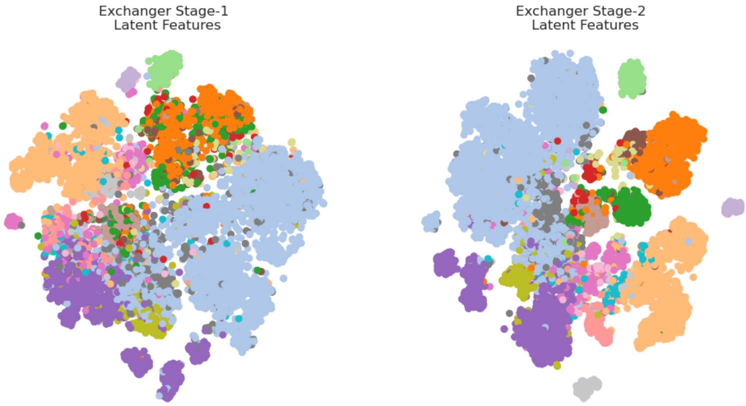

Figure 3: t-SNE [14] visualisations of latent features from stage-1 and stage-2 of Exchanger.

in the output of stage-2 compared to that of stage-1, indicating a hierarchical clustering procedure enabled by increasing the depth of Exchanger. Additionally, it is noticeable that the multi-mode nature inherent in crop type recognition renders the traditional way in NLP of prepending the input sequence with a single class token less effective.

6 Domain Generalization for SITS

In this section, we further present results of the Exchanger[2-stages w/ 8 tokens] evaluated on TimeMatch dataset [9] which is comprised of SITS from four different tiles: 33UVP (Austria), 32VNH (Denmark), 30TXT (mid- west France), and 31TCJ (southern France). We follow the naming convention adopted in [9] to refer to these four Sentinel-2 tiles as AT1, DK1, FR1, and FR2, respectively, and the leave-one-region-out evaluation protocol where one Sentinel-2 tile is held out for testing and the remaining three tiles are used for training. In addition to the specifically-curated dataset for evaluating spatial generalization capability of crop classifiers, authors in [9] proposed to use thermal positional encoding (TPE) to combat temporal shifts across different geographical locations where Growing Degree Days (GDD) have been used to replace calendar time. We directly use the TPE method proposed in

[9] to modify the positional encoding component in Exchanger. Based on our empirical observations, it is favourable to set the dimension of positional embeddings to a relatively small number for better generalization performance, indicating the sensitivity to resolutions of frequencies in sine/cosine functions. As seen in Tab.1, our proposed model trained only for 20 epochs can achieve results comparable to those of PSE+LTAE[4] trained for 100 epochs in the original setup. But the highly-specialized architecture PSE+LTAE[4] still has demonstrated superiority to our model, which we leave as a future direction for improvement.

		AT1	DK1	FR1	FR2	Avg.
PSE+LTAE[4]	TPE-Fourier	84.7	79.0	77.3	80.0	80.3
	TPE-Recurrent	**86.5**	**80.3**	**86.0**	**80.5**	**83.3**
Exchanger	TPE-Fourier	84.1	77.8	84.2	77.6	80.9
	TPE-Recurrent	82.9	80.1	81.2	76.4	80.2

Table 1: Leave-one-region-out spatial generalization results (macro F1 score).

7 Qualitative Results

Figure 4: Qualitative comparison. We randomly sample 4 SITS sample from PASTIS Fold-1 validation dataset and present the panoptic prediction results from U-TAE+PaPs, Exchanger+Unet+PaPs, and Exchanger+Mask2Former. Please note the artefacts in the last column result from stitching 64×64 predictions to 128×128.

In this section, we present a qualitative comparison between previous SOTA model U-TAE + PaPs, Exchanger+Unet+PaPs and the first universal SITS segmentation architecture Exchanger + Mask2Former as a result of concatenating Exchanger as the temporal encoder with the recently proposed universal natural image segmentation framework Mask2Former[2]. As shown in Fig.4, U-TAE+PaPs can retrieve crop parcels almost as the same number as that of Exchanger+PaPs but is more prone to error predictions, which indicates that the weaker representation learning capability of U-TAE. Coupling Exchanger with a more powerful segmentation architecture Mask2Former[2], the panoptic prediction quality is significantly improved in terms of crop type recognition accuracy and crop shape prediction consistent with the SQ and RQ metrics reported in the main body.

8 More Qualitative Visualisations from Exchanger+Mask2Former

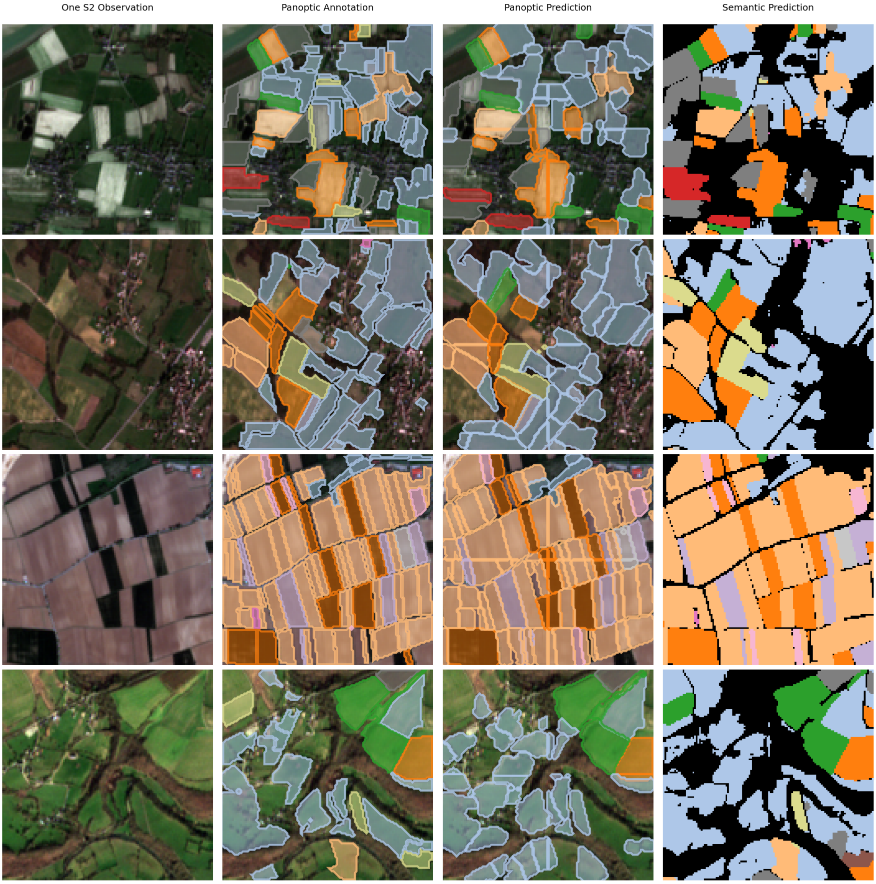

Figure 5: Qualitative Results from predictions of Exchanger+Mask2Former. Please note the segmentation & panoptic segmentation models are separately trained.

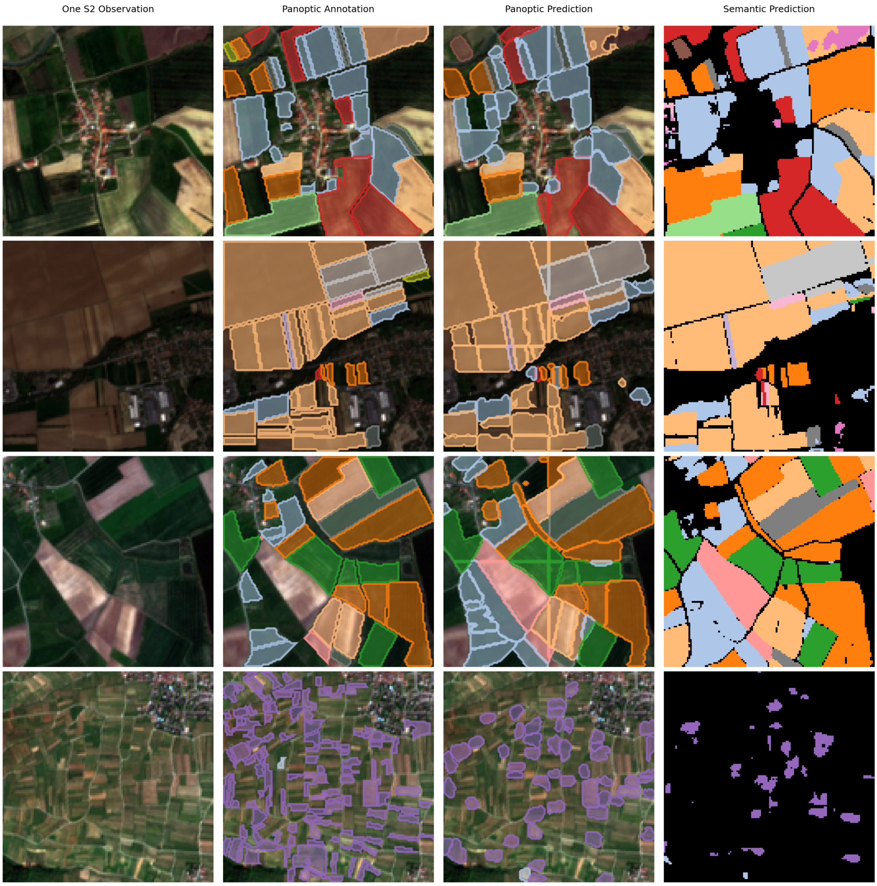

Figure 6: Qualitative Results from predictions of Exchanger+Mask2Former. Please note the segmentation & panoptic segmentation models are separately trained.

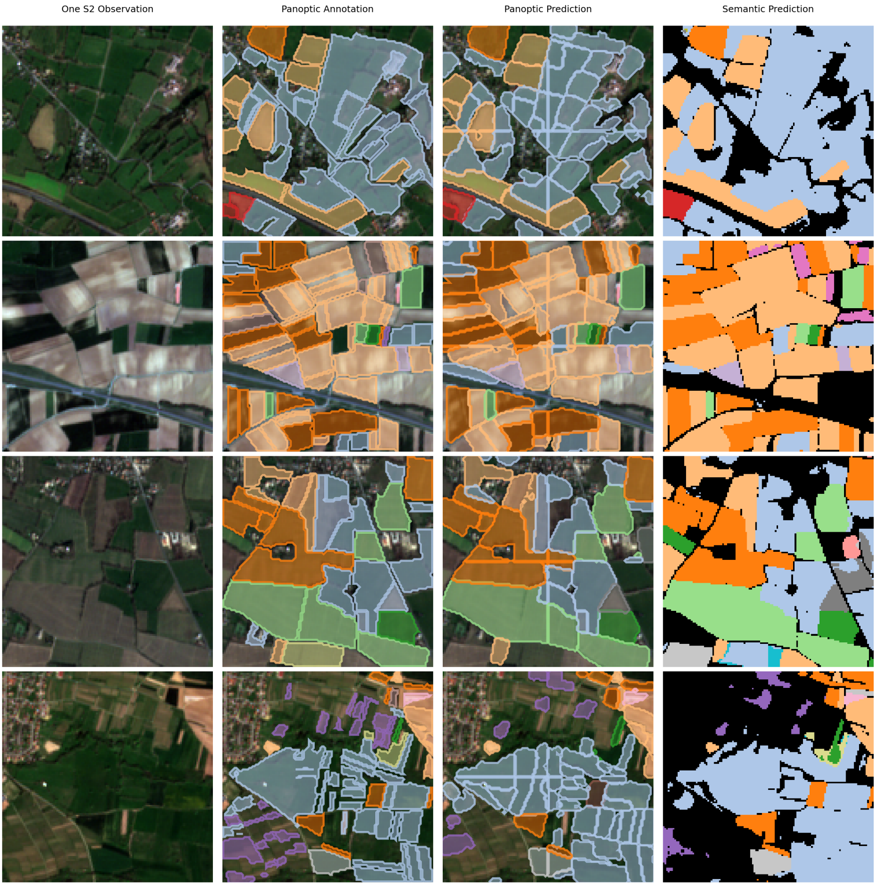

Figure 7: Qualitative Results from predictions of Exchanger+Mask2Former. Please note the segmentation & panoptic segmentation models are separately trained.

References

[1] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14496–14506, 2023.

[2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Vivien Sainte Fare Garnot and Loic Landrieu. Lightweight temporal self-attention for classifying satellite images time series. In *Advanced Analytics and Learning on Temporal Data: 5th ECML PKDD Workshop, AALTD 2020, Ghent, Belgium, September 18, 2020, Revised Selected Papers 6*, pages 171–181. Springer, 2020.

[5] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4872–4881, 2021.

[6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[8] Xu Ma, Yuqian Zhou, Huan Wang, Can Qin, Bin Sun, Chang Liu, and Yun Fu. Image as set of points. In *The Eleventh International Conference on Learning Representations*, 2023.

[9] Joachim Nyborg, Charlotte Pelletier, and Ira Assent. Generalized classification of satellite image time series with thermal positional encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1392–1402, 2022.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[11] Mert Bulent Sariyildiz, Yannis Kalantidis, Karteek Alahari, and Diane Larlus. No reason for no supervision: Improved generalization in supervised models. In *ICLR 2023-International Conference on Learning Representations*, pages 1–26, 2023.

[12] Teppei Suzuki. Clustering as attention: Unified image segmentation with hierarchical clustering. *arXiv preprint arXiv:2205.09949*, 2022.

[13] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

[14] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[15] Yizhou Wang, Shixiang Tang, Feng Zhu, Lei Bai, Rui Zhao, Donglian Qi, and Wanli Ouyang. Revisiting the transferability of supervised pretraining: an mlp perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9183–9193, 2022.

[16] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022.

[17] Chenhongyi Yang, Jiarui Xu, Shalini De Mello, Elliot J Crowley, and Xiaolong Wang. Gpvit: A high resolution non-hierarchical vision transformer with group propagation. *arXiv preprint arXiv:2212.06795*, 2022.

[18] Qihang Yu, Huiyu Wang, Dahun Kim, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Cmt-deeplab: Clustering mask transformers for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2560–2570, 2022.