

Diminishing Returns in Self-Supervised Learning

Oli Bridge* Huey Sun* Botond Branyicskai-Nagy* Charles D’Ornano*
Shomit Basu*

University College London, London UK ucabhss@ucl.ac.uk
<https://www.ucl.ac.uk/computer-science/>

**Equal contribution*

Abstract. Transformer-based architectures have become a dominant paradigm in vision and language, but their success is often attributed to large model capacity and massive training data. In this work, we examine how self-supervised pre-training, intermediate fine-tuning, and downstream fine-tuning interact in a low-capacity regime, using a 5M-parameter Vision Transformer for semantic segmentation. Across multiple data scales, we find that masked image modeling pre-training and downstream fine-tuning reliably improve performance, but with clear diminishing returns as supervision increases. In contrast, inserting an intermediate classification fine-tuning stage consistently degrades downstream performance, with the largest drops occurring precisely where pre-training is most effective. Through an analysis of patch-level representation geometry, we show that classification-based intermediate supervision actively interferes with representations learned during pre-training by collapsing spatial structure critical for dense prediction. These results indicate that, in small models, the geometry of supervision matters more than the number of training stages: misaligned intermediate objectives can negate the benefits of pre-training rather than amplify them.

1 Introduction

1.1 Background

Deep learning has driven major advances in computer vision, from image classification and object detection to semantic segmentation. However, many practical settings still face a scarcity of high-quality labeled data. Self-supervised learning (SSL) addresses this by pre-training models on large collections of unlabeled images and then fine-tuning them for a supervised downstream task, using pre-training to learn transferable representations.

Vision Transformers (ViTs) are a compelling architecture for SSL because they scale effectively with data and compute, representing images as sequences of patch tokens processed by self-attention [8]. While prior work has demonstrated the benefits of SSL for Vision Transformers at scale [5], it remains unclear how these methods behave when constrained to small models and limited data. In this work, we use a small 5M-parameter ViT (ViNy) to systematically examine how



Fig. 1. Training pipeline for ViNy. The optional intermediate fine-tuning stage uses global (classification-based) or spatial (patch-aligned) supervision to isolate how the geometry of supervision affects transfer in low-capacity Vision Transformers.

the amount of pre-training and fine-tuning data affects downstream semantic segmentation performance.

For pre-training, we use masked image modeling (MIM), where a model reconstructs masked portions of an input image [4]. We adopt SimMIM [12], which masks image patches and predicts missing pixel values with a lightweight decoder, and pre-train on ImageNet-1K [11]. We also study intermediate fine-tuning [10], an additional supervised training stage performed before downstream fine-tuning that is commonly used to improve transfer in large models, but whose effect on dense prediction in low-capacity Vision Transformers remains poorly understood.

Overall, we evaluate how three stages—SimMIM pre-training, intermediate fine-tuning, and downstream segmentation fine-tuning—interact across data regimes, and when intermediate supervision helps versus harms transfer (Figure 1). We show that classification-based intermediate fine-tuning actively destroys spatial structure in small Vision Transformers, negating the benefits of self-supervised pre-training.

1.2 Related Literature

Intermediate fine-tuning has been extensively studied in natural language processing as a mechanism for improving transfer in pre-trained models. In NLP, self-supervised pre-training underlies models such as BERT [6], while supervised intermediate fine-tuning on data-rich tasks [10] was initially thought to improve downstream performance by transferring general skills.

Subsequent work has shown that this interpretation is incomplete. Chang and Lu [3] demonstrate that the effectiveness of intermediate fine-tuning depends on factors such as dataset size, task format, and alignment between training stages, and that intermediate objectives can influence downstream performance even when they are not semantically related. These findings raise broader questions about how intermediate supervision reshapes learned representations.

In computer vision, pre-training and intermediate fine-tuning have also been widely explored. Masked modeling has been successfully adapted to images through methods such as BEiT [2] and SimMIM [12], and intermediate fine-tuning is commonly applied to improve downstream classification and dense



Fig. 2. Example SimMIM reconstructions on ImageNet-1K. Despite its small capacity (5M parameters), ViNy learns non-trivial patch-level structure during masked image modeling, providing a meaningful initialization for downstream tasks.

prediction performance. Prior work typically uses intermediate objectives that are closely aligned with the final task, and has shown that effective pre-training does not require massive datasets, particularly in limited-capacity regimes [7].

In contrast to prior vision work that emphasizes aligned intermediate objectives, we study intermediate fine-tuning under intentional task mismatch in a low-capacity Vision Transformer. Rather than proposing a new training strategy, we investigate when intermediate fine-tuning fails, and use simple representation diagnostics to characterize how misaligned supervision alters patch-level feature structure. This shifts attention from whether intermediate fine-tuning improves performance to when and why it can be detrimental in small-model, low-compute regimes.

2 Methodology

2.1 ViNy

All of our experiments use a small ViT architecture with an input image size of 128, a patch size of 16, an embedding dimension of 128, a depth of 12, 8 attention heads, and an MLP dimension of 512. This configuration, which we refer to as ViNy (for "tiny ViT"), contains 4.8M parameters, representing an $18\times$ reduction compared to ViT-Base's 86M parameters.

2.2 Pre-training with SimMIM

We perform unsupervised pre-training using SimMIM [12], which randomly masks image patches and trains the model to reconstruct masked pixels using an L1 loss. We adopt a masking ratio of 0.5 and pre-train on ImageNet-1K [11], varying the number of pre-training examples between 0, 45k, 100k, and 200k. Optimization details are provided in the appendix in Table A1.

2.3 Intermediate Fine-tuning

Unlike prior work that applies intermediate objectives aligned with the final task, we intentionally introduce a mismatch between intermediate and downstream supervision. Specifically, we perform intermediate fine-tuning on the Intel Image Classification dataset [1], which contains approximately 14,000 images across six natural scene categories (e.g., buildings, forests, mountains). During this stage, ViNy is equipped with a 128×6 classification head, which is discarded before downstream segmentation fine-tuning. This design isolates the effect of global classification supervision on representations later used for dense prediction.

While a more closely aligned intermediate task would likely improve downstream accuracy, our goal is not to optimize performance but to characterize how misaligned intermediate supervision reshapes representations and affects transfer.

2.4 Downstream Segmentation Fine-tuning

The downstream task is semantic segmentation on the Oxford-IIIT Pet dataset [9], using tri-map labels for foreground, background, and ambiguous pixels. We vary the number of training examples between 250 and 6000 and evaluate on 1000 held-out test images using a fixed random seed. ViNy is equipped with a patch-level prediction head mapping each patch embedding to a $16 \times 16 \times 3$ output. We report pixel accuracy and mean intersection-over-union (mIoU). Example segmentation outputs are shown in Appendix Figure A1. Code to reproduce all experiments is at <https://github.com/itshuey/masked-image-modelling>.

3 Experiments

First, we conduct an experiment to determine how the baseline ViNy model performs as we increase the number of fine-tuning examples, evaluating accuracy and mIoU on 1000 held-out test examples (Table 1).

Table 1. Baseline ViNy performance over 3 runs by number of fine-tuning examples

# Finetune	Accuracy	mIoU
250	$70.33 \pm 0.89\%$	$43.73 \pm 1.02\%$
500	$73.71 \pm 0.08\%$	$47.41 \pm 0.21\%$
1000	$77.81 \pm 0.56\%$	$53.28 \pm 0.43\%$
2000	$80.03 \pm 0.42\%$	$56.87 \pm 0.51\%$
4000	$84.16 \pm 0.06\%$	$62.99 \pm 0.16\%$
6000	$85.60 \pm 0.03\%$	$65.07 \pm 0.01\%$

We then repeat this experiment for ViNy models pre-trained using SimMIM with a varying number of training examples from ImageNet 1-K. The results are visualized in Figure 3 and listed in Table A2 in the appendix.

Fig. 3. Downstream segmentation mIoU as a function of pre-training scale and downstream fine-tuning data.

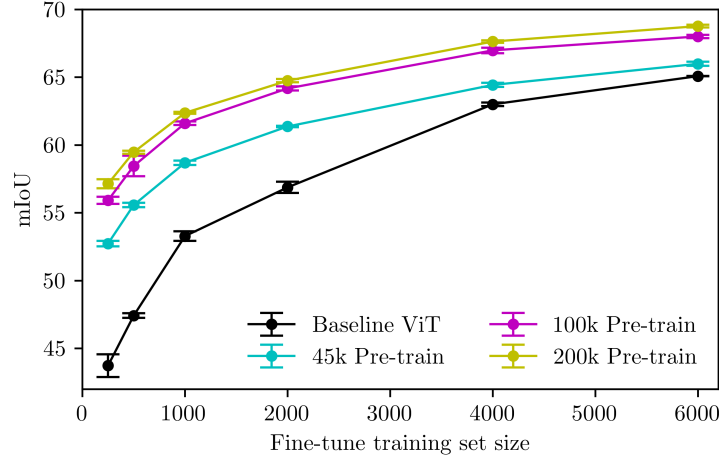
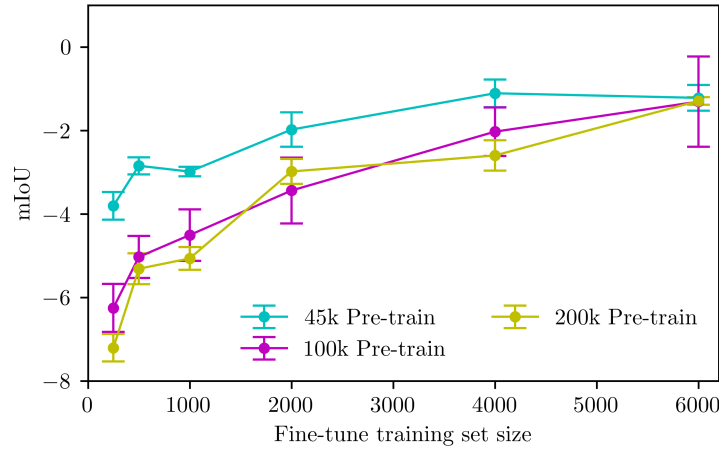


Fig. 4. Change in downstream mIoU induced by adding an intermediate classification fine-tuning stage. All models receive identical intermediate supervision on Intel Image Classification between pre-training and downstream segmentation fine-tuning.



We then repeat this experiment with the inclusion of an intermediate fine-tuning stage. The impact of this additional supervision is visualized in Figure 4 and the data is listed in Table A3 in the appendix. As intermediate fine-tuning worsens performance, we show the paired results in terms of negative mIoU. All experiments are averaged over three runs with fixed data splits to control for variance.

Table 2. Patch-level representation diagnostics across training regimes. Patch Probe Acc measures linear separability of patch features for segmentation classes.

Model	Within-Image Variance	Across-Image Variance	Patch Probe Acc (%)
Pretrained (SimMIM)	0.10	0.02	79.5
CLS-Intermediate (Intel)	0.18	0.11	76.1
CLS-Intermediate (Pets)	0.58	0.29	78.6
Spatial Intermediate	0.32	0.06	93.3

4 Results

4.1 Performance effects of pre-training and intermediate supervision

Pre-training substantially improves downstream segmentation performance in low-data regimes, yielding gains of up to approximately 13 mIoU points, but exhibits clear diminishing returns as the amount of downstream supervision increases (Figure 3). Increasing the scale of pre-training beyond 100k ImageNet examples produces smaller marginal improvements, which for a model as small as ViNy (5M parameters) likely reflects a capacity bottleneck rather than insufficient data.

In contrast, adding an intermediate classification fine-tuning stage consistently degrades downstream segmentation performance across all configurations (Figure 4). The degradation is largest precisely in regimes where SimMIM pre-training is most effective, indicating that intermediate supervision interferes with—rather than complements—representations learned during self-supervised pre-training.

4.2 Intermediate fine-tuning alters spatial representation geometry

To better understand this degradation, we analyze how different types of supervision reshape the spatial geometry of patch-level representations. Because semantic segmentation depends on preserving fine-grained spatial structure, effective representations should be locally coherent within regions while still allowing sharp discontinuities at object boundaries. Accordingly, patch embeddings should exhibit higher similarity for nearby patches on average, with similarity decreasing as spatial distance increases, while permitting non-local similarity between semantically related regions.

To isolate the role of supervision geometry, we include *spatial intermediate* supervision as a diagnostic control. In this setting, the model is trained on the same dataset as the classification-based intermediate task, but with patch-aligned, spatially local targets rather than a global classification objective, allowing us to attribute observed differences to supervision geometry rather than dataset shift or supervision strength.

Fig. 5. Patch similarity as a function of spatial distance on the ViT patch grid. Curves show cosine similarity between patch embeddings at increasing Manhattan distances.

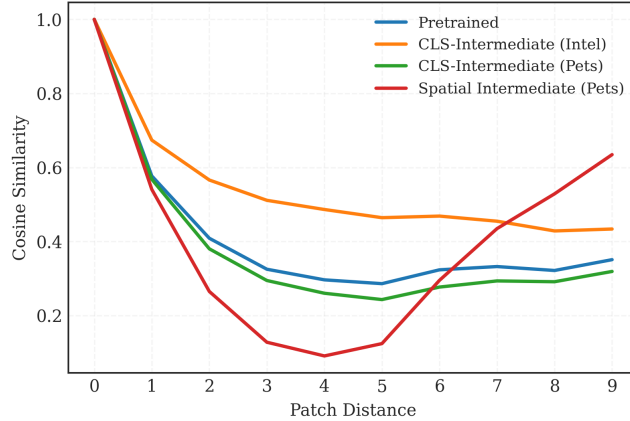


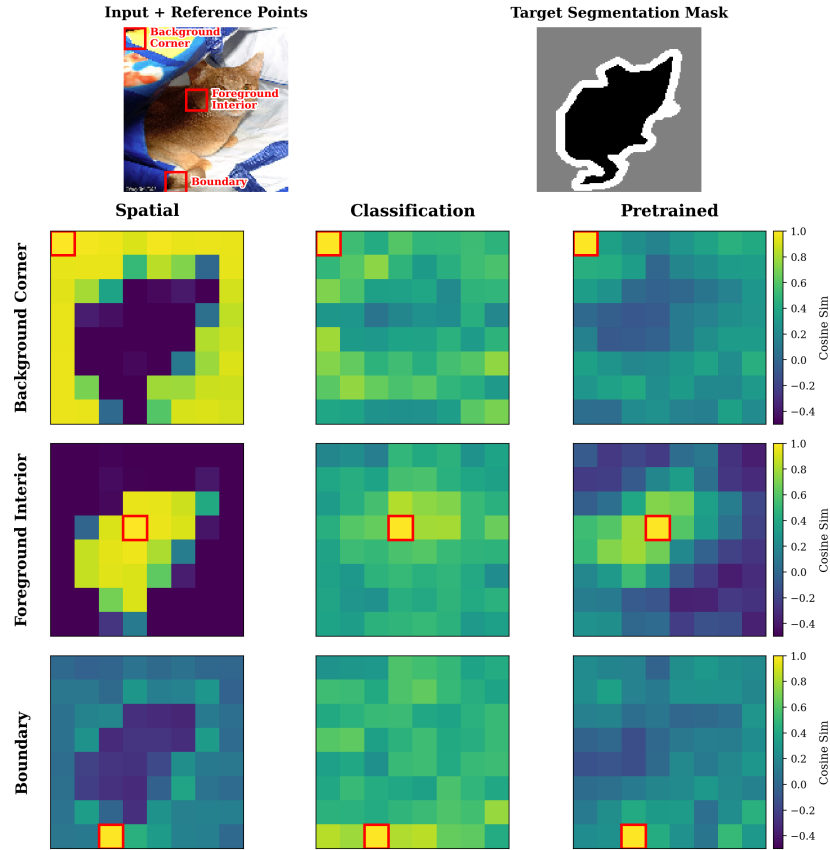
Figure 5 plots cosine similarity between patch tokens as a function of spatial distance on the ViT patch grid. Across most regimes—including SimMIM pre-training and classification-based intermediate fine-tuning on Intel—similarity declines gradually with distance. Classification-based intermediate fine-tuning on the downstream dataset (CLS-Pets) increases overall similarity but does not substantially alter the shape of the similarity-distance curve, suggesting a dataset-specific effect rather than a fundamental geometric change.

At first glance, the similar distance-decay profiles for SimMIM pre-training and CLS-Intel raise a puzzle: if the curves are nearly identical, why does intermediate fine-tuning degrade downstream segmentation performance? The distinction lies not in the average similarity profile, but in the structure of similarity across individual patches.

Figure 6 makes this concrete: each heatmap visualizes cosine similarity between a reference patch (red square) and all other patches in the image. Representations learned with SimMIM pre-training exhibit weak but discernible spatial structure, with coherent regions emerging amidst background noise. Classification-based intermediate fine-tuning preserves the overall similarity scale but substantially increases spatial noise, obscuring region-level organization even when average distance-dependent trends remain unchanged. The resulting segmentation predictions exhibit spatially unstructured errors (Appendix Figure A2).

In contrast, spatially aligned intermediate supervision induces a qualitatively different representational structure. Nearby patches within the same region exhibit high similarity (local coherence), while distant patches corresponding to the same semantic class—most often background—also show elevated similarity (semantic grouping). Together, these complementary structures produce the characteristic U-shaped profile in Figure 5, where similarity is high at both short and long distances but drops at intermediate scales. This reflects semantically

Fig. 6. Emergent patch-level similarity structure under different training regimes. Each heatmap shows cosine similarity between a reference patch (red square) and all other patches in the image. Classification-based intermediate fine-tuning produces spatially diffuse, globally similar representations, while spatially aligned supervision preserves localized structure consistent with dense prediction.



organized patch representations suitable for dense prediction, rather than the noisy, unstructured patterns produced by classification-based supervision.

These geometric differences translate directly into downstream utility. As shown in Table 2, spatial intermediate supervision yields highly linearly separable patch features, achieving substantially higher probe accuracy than either pre-training alone or classification-based intermediate fine-tuning. In contrast, classification-based supervision increases variance without improving the usability of patch representations for dense prediction.

Together, these results show that in ViNy, intermediate fine-tuning reshapes patch-level representations in ways that directly affect downstream segmentation performance. Spatially aligned supervision sharpens and organizes patch

representations, while classification-based supervision primarily injects noise and weakens the spatial structure learned during pre-training.

5 Discussion

Our results suggest that, in low-capacity Vision Transformers, the benefits of additional training stages depend critically on the geometry of supervision rather than the number of stages alone. While self-supervised pre-training provides a strong initialization for dense prediction, intermediate classification fine-tuning can systematically disrupt spatial structure and negate these gains.

A plausible explanation for this behavior lies in the geometry of the classification objective itself. In Vision Transformers, classification loss is mediated through a global [CLS] token, encouraging patch tokens to contribute to image-level discrimination rather than encode spatially local features. This incentive structure increases variance across images while weakening spatial organization within them (Table 2), producing the spatially unstructured representations observed under classification-based intermediate fine-tuning. When representational capacity is limited and features are easily overwritten, this global pressure overwrites spatial structure learned during self-supervised pre-training rather than complementing it.

These findings suggest a practical guideline for low-compute regimes: prioritize downstream supervision, apply self-supervised pre-training where possible, and use intermediate fine-tuning only when its supervision structure aligns with the downstream task. In small models, task alignment matters more than the number of training stages, and indiscriminate intermediate fine-tuning can undo the benefits of pre-training rather than amplify them.

References

1. Intel image classification dataset, <https://www.kaggle.com/puneet6060/intel-image-classification>
2. Bao, H., Dong, L., Piao, S., Wei, F.: BEit: BERT pre-training of image transformers. In: International Conference on Learning Representations (2022), <https://openreview.net/forum?id=p-BhZSz59o4>
3. Chang, T.Y., Lu, C.J.: Rethinking why intermediate-task fine-tuning works. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 706–713. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021). <https://doi.org/10.18653/v1/2021.findings-emnlp.61>, <https://aclanthology.org/2021.findings-emnlp.61>
4. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1691–1703. PMLR (13–18 Jul 2020), <https://proceedings.mlr.press/v119/chen20s.html>
5. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9620–9629 (2021). <https://doi.org/10.1109/ICCV48922.2021.00950>
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). <https://doi.org/10.18653/v1/N19-1423>, <https://aclanthology.org/N19-1423>
7. El-Nouby, A., Izacard, G., Touvron, H., Laptev, I., Jégou, H., Grave, E.: Are large-scale datasets necessary for self-supervised pre-training? ArXiv **abs/2112.10740** (2021), <https://api.semanticscholar.org/CorpusID:245334705>
8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15979–15988 (2022). <https://doi.org/10.1109/CVPR52688.2022.01553>
9. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3498–3505 (2012). <https://doi.org/10.1109/CVPR.2012.6248092>
10. Phang, J., Févry, T., Bowman, S.R.: Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. CoRR **abs/1811.01088** (2018), <http://arxiv.org/abs/1811.01088>
11. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115** (09 2014). <https://doi.org/10.1007/s11263-015-0816-y>
12. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: a simple framework for masked image modeling. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9643–9653 (2022). <https://doi.org/10.1109/CVPR52688.2022.00943>

6 Appendix

Table A1. Training Configurations for the Optimizer and Learning Rate Scheduler

Phase	Epochs	AdamW Params		MultiStepLR Params	
		Learning Rate	Weight Decay	Gamma	Milestones
Pre-training	100	1×10^{-4}	0.05	0.1	{50, 85}
Intermediate	200	8×10^{-4}	0	0.1	{180, 190}
Fine-tuning	100	2×10^{-3}	1×10^{-4}	0.5	{70, 90, 95}

Fig. A1. A semantic segmentation example on the Oxford III Pets data set. The left image is the input, the center image is the ground truth, and the right image is our model’s prediction



Fig. A2. Example segmentation failures after classification-based intermediate fine-tuning. Errors are spatially diffuse and unstructured, consistent with the loss of spatial locality observed in patch-level similarity analyses.

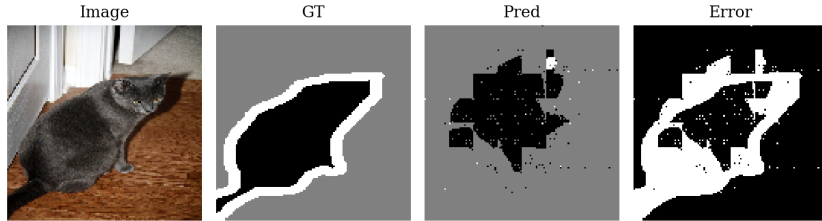


Table A2. Pre-trained ViNy test accuracy and mIoU across 3 runs by number of pre-training examples and number of fine-tuning examples

# Pre-train	# Fine-tune	Accuracy	mIoU
45k	250	77.44 \pm 0.06%	52.73 \pm 0.25%
	500	79.67 \pm 0.20%	55.57 \pm 0.19%
	1000	81.92 \pm 0.20%	58.68 \pm 0.20%
	2000	83.38 \pm 0.03%	61.36 \pm 0.05%
	4000	85.41 \pm 0.13%	64.42 \pm 0.19%
	6000	86.40 \pm 0.11%	65.98 \pm 0.19%
100k	250	79.67 \pm 0.35%	55.91 \pm 0.32%
	500	81.62 \pm 0.62%	58.44 \pm 0.93%
	1000	83.60 \pm 0.14%	61.59 \pm 0.16%
	2000	84.99 \pm 0.18%	64.17 \pm 0.18%
	4000	86.81 \pm 0.12%	66.97 \pm 0.25%
	6000	87.53 \pm 0.08%	68.00 \pm 0.14%
200k	250	80.49 \pm 0.18%	57.14 \pm 0.42%
	500	82.31 \pm 0.04%	59.47 \pm 0.13%
	1000	84.17 \pm 0.15%	62.37 \pm 0.08%
	2000	85.38 \pm 0.11%	64.74 \pm 0.14%
	4000	87.23 \pm 0.06%	67.62 \pm 0.12%
	6000	87.93 \pm 0.04%	68.76 \pm 0.13%

Table A3. Pre-trained and intermediate fine-tuned ViNy test accuracy and mIoU across 3 runs by number of pre-training examples and number of fine-tuning examples

# Pre-train	# Fine-tune	Accuracy	mIoU
45k	250	74.65 \pm 0.97%	48.92 \pm 0.66%
	500	77.86 \pm 0.32%	52.72 \pm 0.39%
	1000	79.80 \pm 0.11%	55.70 \pm 0.13%
	2000	82.08 \pm 0.57%	59.38 \pm 0.88%
	4000	84.86 \pm 0.28%	63.31 \pm 0.68%
	6000	85.70 \pm 0.42%	64.76 \pm 0.62%
100k	250	75.72 \pm 0.49%	50.20 \pm 0.29%
	500	78.59 \pm 0.25%	53.85 \pm 0.44%
	1000	81.13 \pm 0.19%	57.80 \pm 0.10%
	2000	83.77 \pm 0.26%	61.62 \pm 0.11%
	4000	86.13 \pm 0.11%	65.29 \pm 0.61%
	6000	87.32 \pm 0.48%	67.79 \pm 0.43%
200k	250	75.83 \pm 0.66%	49.93 \pm 0.56%
	500	78.49 \pm 0.69%	54.15 \pm 0.77%
	1000	80.96 \pm 0.26%	57.30 \pm 0.57%
	2000	83.96 \pm 0.40%	61.75 \pm 0.62%
	4000	85.88 \pm 0.48%	65.02 \pm 0.77%
	6000	87.18 \pm 0.14%	67.47 \pm 0.15%