

SELF-SUPERVISED REPRESENTATION LEARNING FOR VISION

by

Ryan Patel

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Graduate Department of Computer Science
University of Toronto

© Copyright 2026 by Ryan Patel

*Unofficial template, not affiliated with or endorsed by the University of Toronto. The author,
committee and results in this sample are fictional. Delete this line before you submit.*

Self-Supervised Representation Learning for Vision

Ryan Patel

Doctor of Philosophy

Graduate Department of Computer Science

University of Toronto

2026

Abstract

Labelled images are expensive, and the representations learned from them transfer poorly to tasks their labels did not anticipate. Self-supervised learning replaces labels with a task the data poses to itself, such as matching two augmented views of one image. This thesis studies which choices in that task decide the quality of the learned representation. Chapter 2 reviews contrastive, distillation and masked-prediction methods. Chapter 3 introduces a view-selection scheme that chooses crops by predicted overlap in content rather than at random. Chapter 4 evaluates it on linear probing, low-shot classification and dense prediction, where it improves ten-shot accuracy by 3.2 points over the strongest random-crop baseline at equal compute.

Acknowledgements

I thank my supervisor for her advice and her calm through three deadlines that moved. I am grateful to my supervisory committee for their careful questions, to my labmates for sharing GPUs and good ideas, and to my family for their support throughout.

Contents

1	Introduction	1
2	Background	2
3	Content-Aware View Selection	3
4	Evaluation	4
5	Conclusion	5
A	Training Details	6

List of Tables

4.1	Linear-probe and ten-shot accuracy (%) at equal compute.	4
-----	--	---

List of Figures

3.1	View selection places a scorer between cropping and the encoder. . .	3
-----	--	---

Chapter 1

Introduction

Deep networks trained on large labelled datasets learn features that transfer surprisingly well, but the labels themselves limit what the features encode. Self-supervised methods learn from unlabelled images by solving a pretext task, and at scale they now match or exceed supervised pre-training on many downstream benchmarks [1, 4].

Most of these methods agree on the broad recipe: create two views of an image, encode both, and train the encoder so the views agree. They disagree on the details, and the details matter. This thesis isolates one of them, how the two views are chosen, and shows that choosing views by their content rather than at random yields better representations for the same compute.

Chapter 2

Background

Given an encoder f_θ and two augmented views x_i, x_j of the same image, contrastive methods minimize a normalized temperature-scaled cross-entropy,

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k \neq i} \exp(\text{sim}(z_i, z_k)/\tau)}, \quad (2.1)$$

where $z = g(f_\theta(x))$ is a projected embedding and the denominator runs over the other views in the batch [1]. Distillation methods drop the negatives and predict one view’s embedding from the other’s [3], while masked autoencoders reconstruct hidden patches of a single view [5].

Chapter 3

Content-Aware View Selection

Random crops often produce views that share little content, such as two patches of sky from a photograph of a boat, and Equation (2.1) then asks the encoder to match unrelated patches. We train a lightweight scorer that predicts the semantic overlap of a candidate crop pair from a low-resolution saliency map, and sample pairs with probability proportional to the predicted overlap, tempered so that some hard pairs remain.

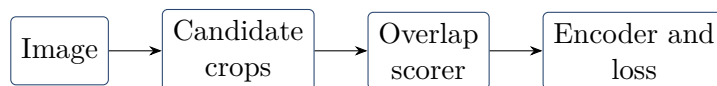


Figure 3.1: View selection places a scorer between cropping and the encoder.

Chapter 4

Evaluation

All methods use a ViT-S/16 encoder [2] trained for 300 epochs on the same images. Table 4.1 reports linear-probe and ten-shot accuracy.

Table 4.1: Linear-probe and ten-shot accuracy (%) at equal compute.

Method	Linear probe	10-shot
Contrastive, random crops	73.1	51.4
Distillation, random crops	74.6	53.0
Distillation, content-aware	75.9	56.2

The gain is largest in the low-shot setting, where the features must separate classes from very few examples, and it persists on semantic segmentation, where the content-aware model improves mean intersection-over-union by 1.1 points.

Chapter 5

Conclusion

How views are chosen is a first-order design decision in self-supervised learning, not a detail. Scoring candidate views by predicted overlap is cheap, composes with existing objectives, and gives better features at equal cost. Extending the scorer to video, where temporal overlap is a natural signal, is the obvious next step.

Appendix A

Training Details

All models were trained with AdamW, a cosine learning-rate schedule with 40 warm-up epochs, and a batch size of 1024 across eight accelerators. The overlap scorer has 0.6 million parameters and adds 3% to the cost of each training step.

Bibliography

- [1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607, 2020.
- [2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- [3] Jean-Bastien Grill, Florian Strub, Florent Altché, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284, 2020.
- [4] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.
- [5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.