



The Role of Entropy and Reconstruction for Multi-View Self-Supervised Learning

Borja Rodríguez-Gálvez, Arno Blaas, Pau Rodríguez, Adam Goliński, Xavier Suau, Jason Ramapuram, Dan Busbridge, Luca Zappella

ICML 2023

Motivation: MVSSL progressing rapidly

arXiv:2104.14294v2 [cs.CV] 24 May 2021

Emerging Properties in Self-Supervised Vision Transformers

Mathilde Caron^{1,2} Hugo Touvron^{1,3} Ishan Misra¹ Hervé Jegou¹
Julien Mairal² Piotr Bojanowski¹ Armand Joulin¹
¹ Facebook AI Research ² Inria³ Sorbonne University

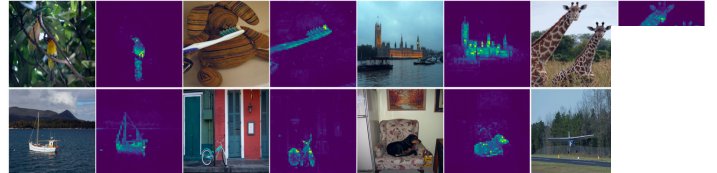


Figure 1: Self-attention from a Vision Transformer with 8 × 8 patches trained with no supervision. We look at the [CLS] token on the heads of the last layer. This token is not attached to any label nor supervision. These maps st automatically learn class-specific features leading to unsupervised object segmentations.

Abstract

In this, video new, stand out, Beyond the architecture ing observers explicit inf image, with VITs, nor a celled k-N with a sma momentum use of sma into a tiny we interpret We show it 80.1% top-

1. Introduction

This paper presents SimCLR, a simple framework for contrastive learning of visual representations. We simplify recently proposed contrastive self-supervised learning algorithms without requiring specialized architectures or a memory bank. In order to understand what enables the contrastive prediction tasks to learn useful representations, we systematically study the major components of our framework. We show that (1) composition of data augmentations plays a critical role in defining effective predictive tasks, (2) introducing a learnable nonlinear transformation between the representation and the contrastive loss substantially improves the quality of the learned representations, and (3) contrastive learning benefits from larger batch sizes and more training steps compared to supervised learning. By combining these findings, we are able to considerably outperform previous methods for self-supervised and semi-supervised learning on ImageNet. A linear classifier trained on self-supervised representations learned by SimCLR achieves 76.5% top-1 accuracy, which is a 7% relative improvement over previous state-of-the-art, matching the performance of a supervised ResNet-50. When fine-tuned on only 1% of the labels, we achieve 83.8% top-5 accuracy, outperforming AlexNet with 100× fewer labels.¹

1. Introduction

Learning effective visual representations without human supervision is a long-standing problem. Most mainstream approaches fall into one of two classes: generative or discriminative. Generative approaches learn to generate or otherwise model pixels in the input space (Hinton et al., 2006; Kingma & Welling, 2013; Goodfellow et al., 2014) not a memory bank (Wu et al., 2018; Tian et al., 2019; He et al., 2019; Misra & van der Maaten, 2019).

In order to understand what enables good contrastive representation learning, we systematically study the major components of our framework and show that:

Abstract

Yonglong Tian MIT CSAIL yonglongt@mit.edu
Dilip Krishnan Google Research dilipkr@google.com
Phillip Isola MIT CSAIL philip@mit.edu

Abstract

Humans view the world through many sensory channels, e.g. the long-wavelength light channel, viewed by the left eye, or the high-frequency vibrations channel, heard by the right ear. Each view is noisy and incomplete, but important factors, such as physics, geometry, and semantics, tend to be shared between all views (e.g., a “dog” can be seen, heard, and felt). We investigate the classic hypothesis that a powerful representation is one that models view-invariant factors. We study this hypothesis under the framework of multi-view contrastive learning, where we learn an approximation that aims to maximize mutual information between different views of the same scene but is otherwise compact. Our approach scales to any number of views, and is view-agnostic. We analyze key properties of the approach that make it work, finding that the contrastive loss outperforms a popular alternative based on cross-view prediction, and that the more views we learn from, the better the resulting representation captures underlying scene semantics. Our approach achieves state-of-the-art results on image and video unsupervised learning benchmarks. Code is released at: <https://github.com/yonglongtian/mvssl>.

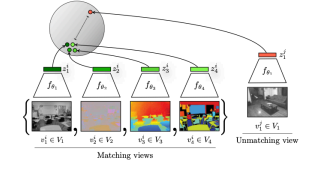


Figure 1: Given a set of sensory views, a deep representation is learned by bringing views of the same scene together in embedding space, while pushing views of different scenes apart. How we show and example of a view dataset (NYU RGBD) [13] and its learned representation. The encoding for each view may be concatenated to form the full representation of a scene.

1. Introduction

Pre-trained convolutional neural network blocks in most computer vision general-purpose features that can be learned on a limited amount of large fully-supervised datasets, however, Stock and Clise that the performance of state-of-the-art is not as good as it seems. However, Stock and Clise that the performance of state-of-the-art is not as good as it seems. However, Stock and Clise that the performance of state-of-the-art is not as good as it seems.

Abstract

Ben Poole Google Research benpoole@google.com
Phillip Isola MIT CSAIL philip@mit.edu

Abstract

Contrastive learning between multiple views of the data has recently achieved state-of-the-art performance in the field of self-supervised representation learning. Despite its success, the influence of different view choices has been less studied. In this paper, we use theoretical and empirical analysis to better understand the importance of view selection, and argue that we should reduce the mutual information (MI) between views while keeping task-relevant information intact. To verify this hypothesis, we devise unsupervised and semi-supervised frameworks that learn effective views by aiming to reduce their MI. We also consider data augmentation as a way to reduce MI, and show that increasing data augmentation indeed leads to decreasing MI and improves downstream classification accuracy. As a by-product, we achieve a new state-of-the-art accuracy on unsupervised pre-training for ImageNet classification (73% top-1 linear readout with a ResNet-50).

1. Introduction

It is commonsense that how you look at an object does not change its identity. Nonetheless, Jorge Luis Borges imagined the alternative. In his short story on *Funes the Memorious*, the titular character becomes bothered that “a dog at three fourteen (seen from the side) should have the same name as the dog at three fifteen (seen from the front)” [8]. The curse of Funes is that he has a perfect memory, and every new way he looks at the world reveals a percept minutely distinct from anything he has seen before. He cannot collate the disparate experiences.

Most of us, fortunately, do not suffer from this curse. We build mental representations of identity that discard nuisance like time of day and viewing angle. The ability to build up view-invariant representations is central to a rich body of research on multi-view learning. These methods seek representations of the world that are invariant to a family of viewing conditions. Currently, a popular paradigm is contrastive multi-view learning, where two views of the same scene are brought together in representation space, and two views of different scenes are pushed apart.

This is a natural and powerful idea but it leaves open an important question: “which viewing conditions should be invariant to?” It’s possible to go too far: if our task is to classify the time of day then we certainly should not use a representation that is invariant to time. Or, like Funes, we could go far enough: representing each specific viewing angle independently would cripple our ability to track a dog as it moves about a scene.

We therefore seek representations with enough invariance to be robust to inconsequential variations but not so much as to discard information required by downstream tasks. In contrastive learning, we achieve a new state-of-the-art accuracy on unsupervised pre-training for ImageNet classification (73% top-1 linear readout with a ResNet-50).

Project page: <https://hobbitlong.github.io/infobits>

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Exploring Simple Siamese Representation Learning

Xinlei Chen Kaiming He
Facebook AI Research (FAIR)

arXiv:2011.00184v2 [cs.CV] 18 Mar 2019

Abstract

Siamese networks have become a common structure for unsupervised visual representation learning. These models maximize the similarity between two augmentations of one image, subject to conditions for avoiding collapsing solutions. In this paper, we report surprising empirical results that simple Siamese networks can learn meaningful representations even without negative sample pairs, (i) negative sample pairs, (ii) momentum encoders. Our experiments show that this collection of solutions can be used for the first time to avoid line collapse.

1. Introduction

Pre-trained convolutional neural network blocks in most computer vision general-purpose features that can be learned on a limited amount of large fully-supervised datasets, however, Stock and Clise that the performance of state-of-the-art is not as good as it seems. However, Stock and Clise that the performance of state-of-the-art is not as good as it seems.

Abstract

Yonglong Tian MIT CSAIL yonglongt@mit.edu
Dilip Krishnan Google Research dilipkr@google.com
Phillip Isola MIT CSAIL philip@mit.edu

Abstract

Humans view the world through many sensory channels, e.g. the long-wavelength light channel, viewed by the left eye, or the high-frequency vibrations channel, heard by the right ear. Each view is noisy and incomplete, but important factors, such as physics, geometry, and semantics, tend to be shared between all views (e.g., a “dog” can be seen, heard, and felt). We investigate the classic hypothesis that a powerful representation is one that models view-invariant factors. We study this hypothesis under the framework of multi-view contrastive learning, where we learn an approximation that aims to maximize mutual information between different views of the same scene but is otherwise compact. Our approach scales to any number of views, and is view-agnostic. We analyze key properties of the approach that make it work, finding that the contrastive loss outperforms a popular alternative based on cross-view prediction, and that the more views we learn from, the better the resulting representation captures underlying scene semantics. Our approach achieves state-of-the-art results on image and video unsupervised learning benchmarks. Code is released at: <https://github.com/yonglongtian/mvssl>.



Figure 1: Given a set of sensory views, a deep representation is learned by bringing views of the same scene together in embedding space, while pushing views of different scenes apart. How we show and example of a view dataset (NYU RGBD) [13] and its learned representation. The encoding for each view may be concatenated to form the full representation of a scene.

1. Introduction

Pre-trained convolutional neural network blocks in most computer vision general-purpose features that can be learned on a limited amount of large fully-supervised datasets, however, Stock and Clise that the performance of state-of-the-art is not as good as it seems. However, Stock and Clise that the performance of state-of-the-art is not as good as it seems.

Abstract

Ben Poole Google Research benpoole@google.com
Phillip Isola MIT CSAIL philip@mit.edu

Abstract

Contrastive learning between multiple views of the data has recently achieved state-of-the-art performance in the field of self-supervised representation learning. Despite its success, the influence of different view choices has been less studied. In this paper, we use theoretical and empirical analysis to better understand the importance of view selection, and argue that we should reduce the mutual information (MI) between views while keeping task-relevant information intact. To verify this hypothesis, we devise unsupervised and semi-supervised frameworks that learn effective views by aiming to reduce their MI. We also consider data augmentation as a way to reduce MI, and show that increasing data augmentation indeed leads to decreasing MI and improves downstream classification accuracy. As a by-product, we achieve a new state-of-the-art accuracy on unsupervised pre-training for ImageNet classification (73% top-1 linear readout with a ResNet-50).

1. Introduction

It is commonsense that how you look at an object does not change its identity. Nonetheless, Jorge Luis Borges imagined the alternative. In his short story on *Funes the Memorious*, the titular character becomes bothered that “a dog at three fourteen (seen from the side) should have the same name as the dog at three fifteen (seen from the front)” [8]. The curse of Funes is that he has a perfect memory, and every new way he looks at the world reveals a percept minutely distinct from anything he has seen before. He cannot collate the disparate experiences.

Most of us, fortunately, do not suffer from this curse. We build mental representations of identity that discard nuisance like time of day and viewing angle. The ability to build up view-invariant representations is central to a rich body of research on multi-view learning. These methods seek representations of the world that are invariant to a family of viewing conditions. Currently, a popular paradigm is contrastive multi-view learning, where two views of the same scene are brought together in representation space, and two views of different scenes are pushed apart.

This is a natural and powerful idea but it leaves open an important question: “which viewing conditions should be invariant to?” It’s possible to go too far: if our task is to classify the time of day then we certainly should not use a representation that is invariant to time. Or, like Funes, we could go far enough: representing each specific viewing angle independently would cripple our ability to track a dog as it moves about a scene.

We therefore seek representations with enough invariance to be robust to inconsequential variations but not so much as to discard information required by downstream tasks. In contrastive learning, we achieve a new state-of-the-art accuracy on unsupervised pre-training for ImageNet classification (73% top-1 linear readout with a ResNet-50).

Project page: <https://hobbitlong.github.io/infobits>

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Unsupervised Learning of Visual Features by Contrasting Cluster Assignments

Mathilde Caron^{1,2} Ishan Misra² Julien Mairal¹
Priya Goyal¹ Piotr Bojanowski² Armand Joulin¹
¹ Inria² Facebook AI Research

arXiv:2006.07733v3 [cs.LG] 10 Sep 2020

Abstract

Unsupervised image representations have significantly reduced the performance of contrastive learning algorithms. In this paper, we propose a new approach to self-supervised image representation learning, referred to as *online* and *target* networks, that interact and learn from each other. From an augmented view of an image, we train the online network to predict the target network representation of the same image under a different augmented view. At the same time, we update the target network with a slow-moving average of the online network. While state-of-the-art methods rely on negative pairs, BYTL achieves a new state of the art without them. BYTL reaches 74.3% top-1 classification accuracy on ImageNet using a linear evaluation with a ResNet-50 architecture and 79.6% with a larger ResNet. We show that BYTL performs on par or better than the current state of the art on both transfer and semi-supervised benchmarks. Our implementation and pretrained models are given on GitHub.¹

1. Introduction

Learning good image representations is a key challenge in computer vision [1, 2, 3] as it allows for efficient training on downstream tasks [4, 5, 6, 7]. Many different training approaches have been proposed to learn such representations, usually relying on visual pretext tasks. Among them, state-of-the-art contrastive methods [8, 9, 10, 11, 12] are trained by reducing the distance between representations of different augmented views of the same image (“positive pairs”), and increasing the distance between representations of augmented views from different images (“negative pairs”). These methods need careful treatment of negative pairs [13] by either relying on large batch sizes [8, 12], memory banks [9] or customized mining strategies [4, 11, 12] to retrieve the negative pairs. In addition, their performance critically depends on the choice of image augmentations [8, 12].

In this paper, we introduce Bootstrap Your Own Latent (BYTL), a new algorithm for self-supervised learning of image representations. BYTL achieves higher performance than state-of-the-art contrastive methods (Step 1) baselines [1].

¹Equal contribution, the order of first authors was randomly selected. <https://github.com/stephanie/stephanie-research/tree/master/bytl>



Figure 1: Performance of BYTL on ImageNet (linear evaluation) using ResNet-50 and our best architecture ResNet-200 (2-x), compared to other unsupervised and supervised methods (Step 1) baselines [1].

Abstract

Yonglong Tian MIT CSAIL yonglongt@mit.edu
Chen Sun Google Research chen.sun@google.com
Ben Poole Google Research benpoole@google.com
Dilip Krishnan Google Research dilipkr@google.com
Cordelia Schmid Google Research cordelia@cs.cmu.edu
Phillip Isola MIT CSAIL philip@mit.edu

Abstract

Contrastive learning between multiple views of the data has recently achieved state-of-the-art performance in the field of self-supervised representation learning. Despite its success, the influence of different view choices has been less studied. In this paper, we use theoretical and empirical analysis to better understand the importance of view selection, and argue that we should reduce the mutual information (MI) between views while keeping task-relevant information intact. To verify this hypothesis, we devise unsupervised and semi-supervised frameworks that learn effective views by aiming to reduce their MI. We also consider data augmentation as a way to reduce MI, and show that increasing data augmentation indeed leads to decreasing MI and improves downstream classification accuracy. As a by-product, we achieve a new state-of-the-art accuracy on unsupervised pre-training for ImageNet classification (73% top-1 linear readout with a ResNet-50).

1. Introduction

It is commonsense that how you look at an object does not change its identity. Nonetheless, Jorge Luis Borges imagined the alternative. In his short story on *Funes the Memorious*, the titular character becomes bothered that “a dog at three fourteen (seen from the side) should have the same name as the dog at three fifteen (seen from the front)” [8]. The curse of Funes is that he has a perfect memory, and every new way he looks at the world reveals a percept minutely distinct from anything he has seen before. He cannot collate the disparate experiences.

Most of us, fortunately, do not suffer from this curse. We build mental representations of identity that discard nuisance like time of day and viewing angle. The ability to build up view-invariant representations is central to a rich body of research on multi-view learning. These methods seek representations of the world that are invariant to a family of viewing conditions. Currently, a popular paradigm is contrastive multi-view learning, where two views of the same scene are brought together in representation space, and two views of different scenes are pushed apart.

This is a natural and powerful idea but it leaves open an important question: “which viewing conditions should be invariant to?” It’s possible to go too far: if our task is to classify the time of day then we certainly should not use a representation that is invariant to time. Or, like Funes, we could go far enough: representing each specific viewing angle independently would cripple our ability to track a dog as it moves about a scene.

We therefore seek representations with enough invariance to be robust to inconsequential variations but not so much as to discard information required by downstream tasks. In contrastive learning, we achieve a new state-of-the-art accuracy on unsupervised pre-training for ImageNet classification (73% top-1 linear readout with a ResNet-50).

Project page: <https://hobbitlong.github.io/infobits>

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

... and many more

Motivation: But poorly understood theoretically

 Lens of Information Theory

 Some contrastive MVSSL methods optimize InfoNCE, a lower bound on the Mutual Information (MI)

 What about the other MVSSL methods?

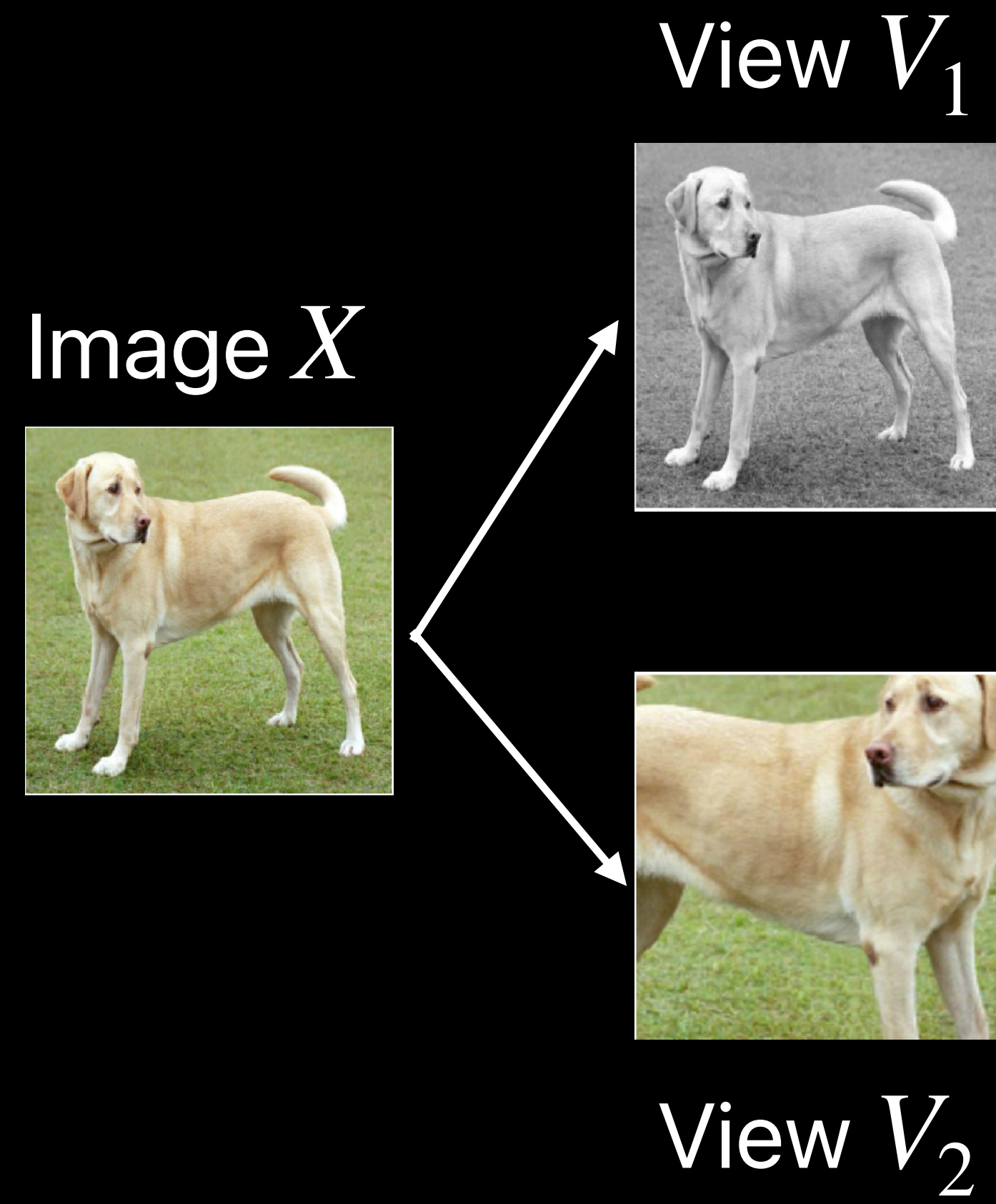
Background: Multi-view self-supervised learning (MVSSL)

Background: Multi-view self-supervised learning (MVSSL)

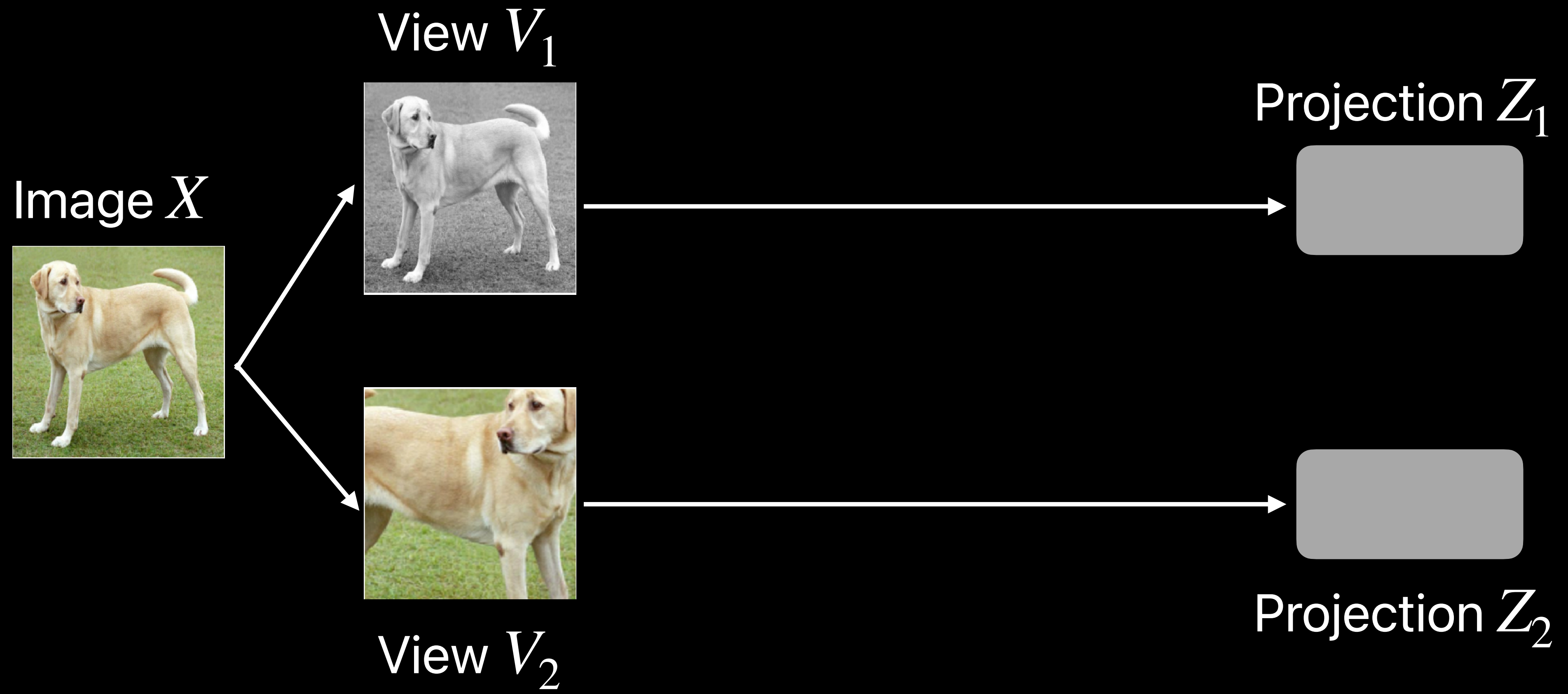
Image X



Background: Multi-view self-supervised learning (MVSSL)



Background: Multi-view self-supervised learning (MVSSL)

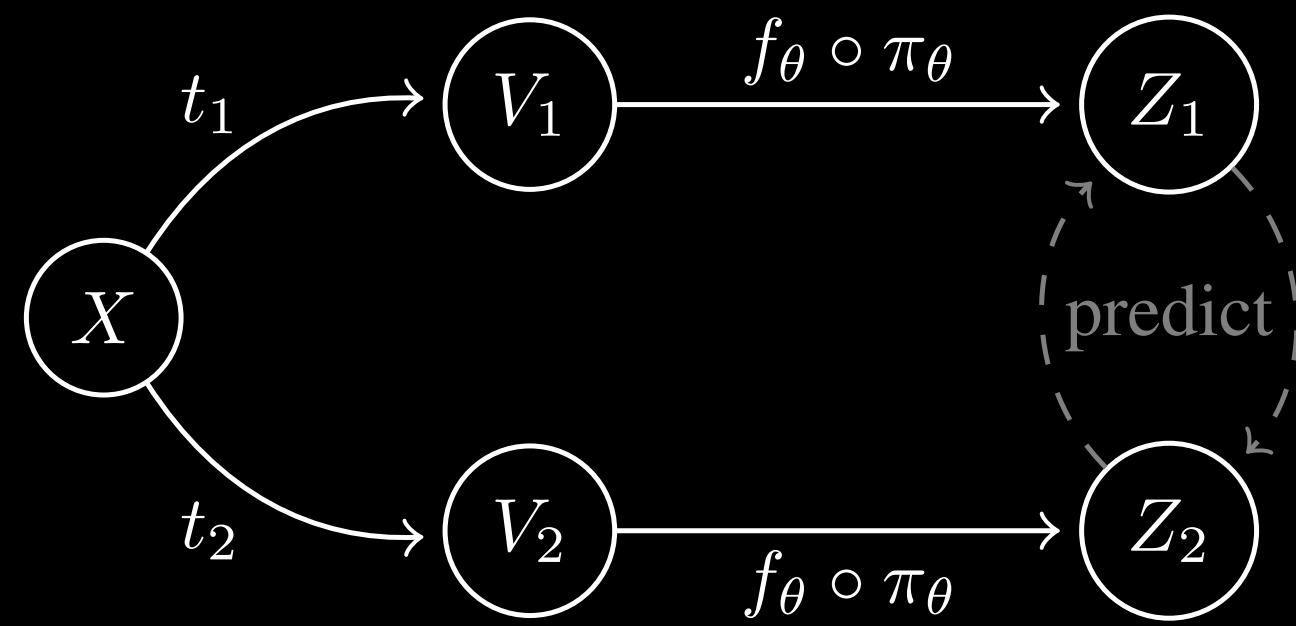


Background: Multi-view self-supervised learning (MVSSL)



Background: MVSSL families

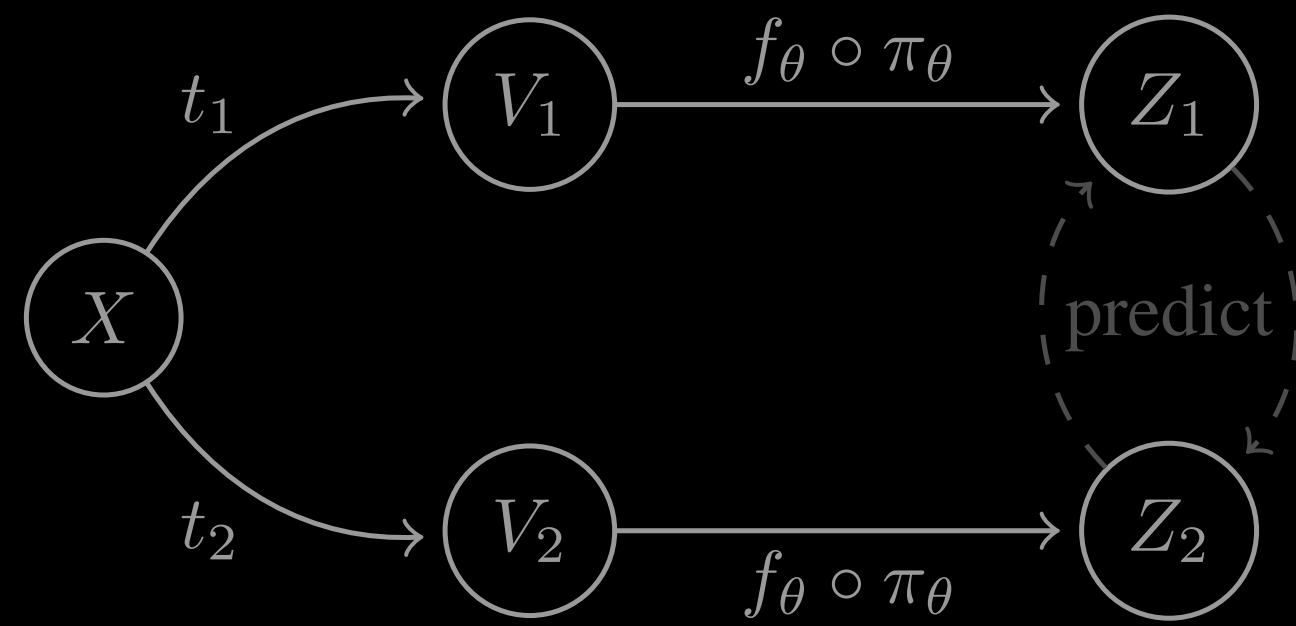
Background: MVSSL families



Contrastive methods

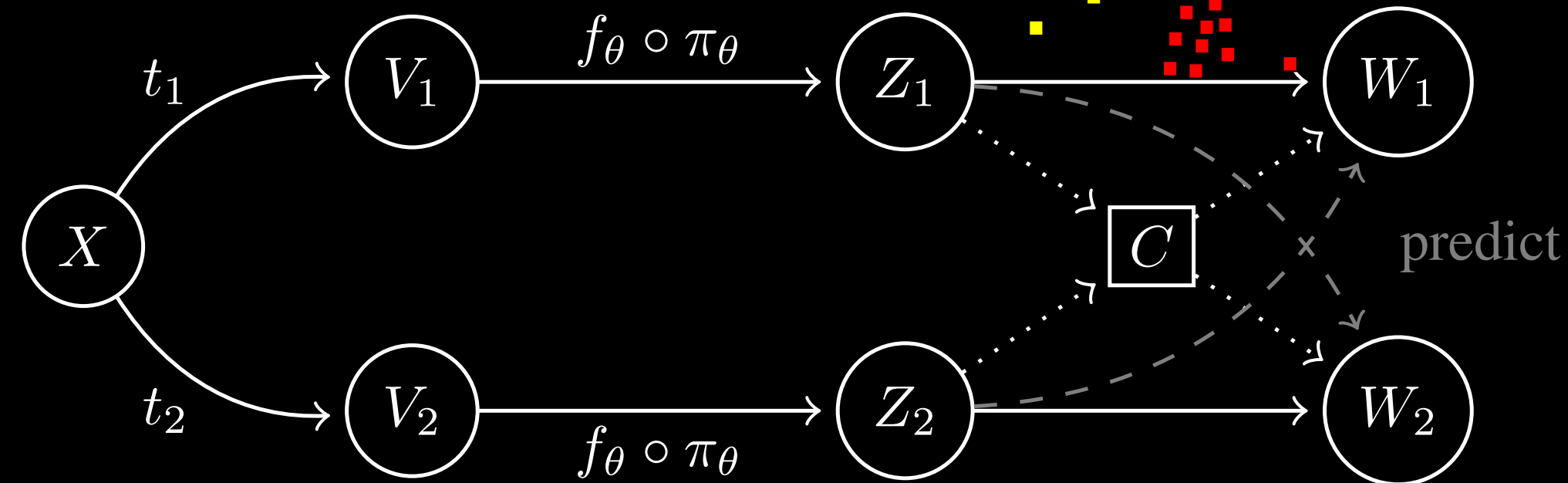
SimCLR, CMC, MoCo

Background: MVSSL families



Contrastive methods

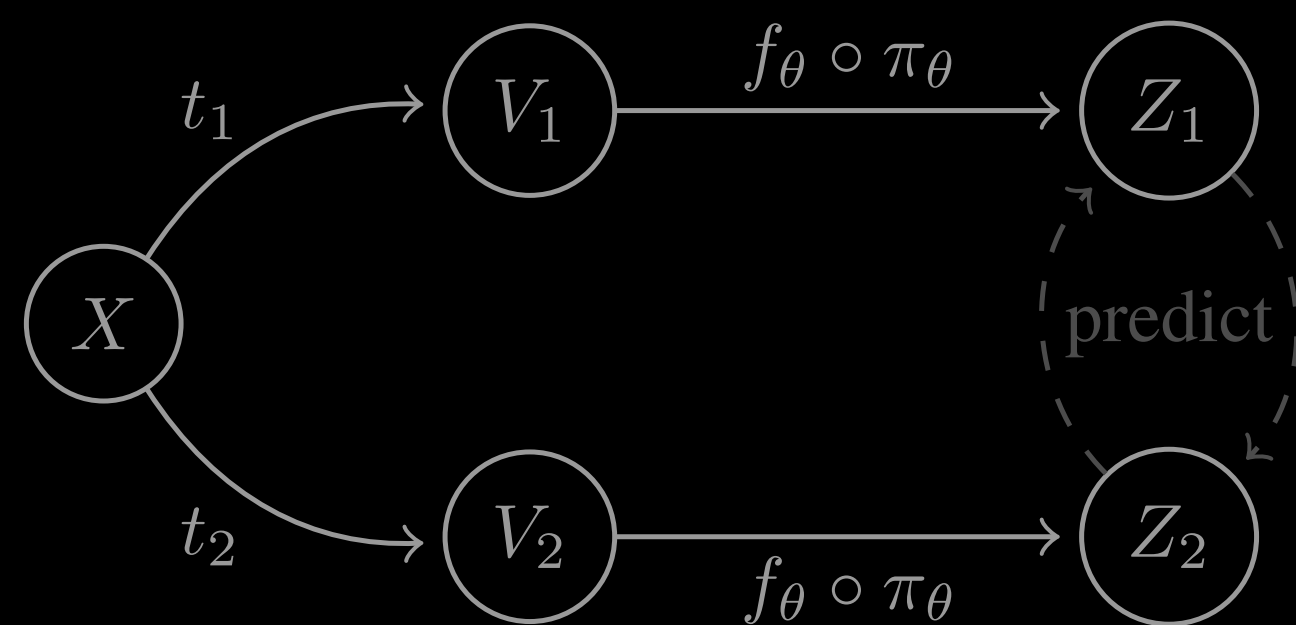
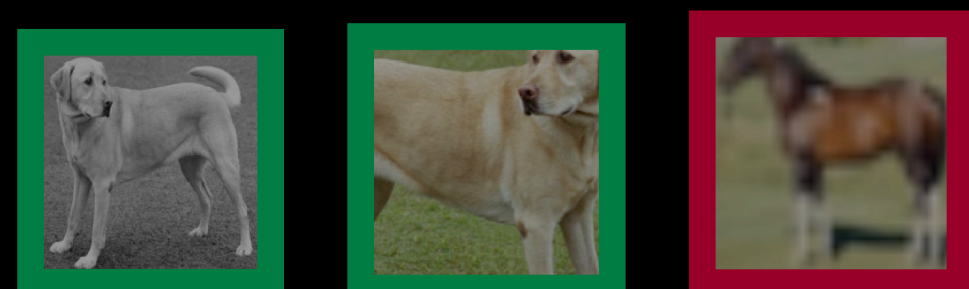
SimCLR, CMC, MoCo



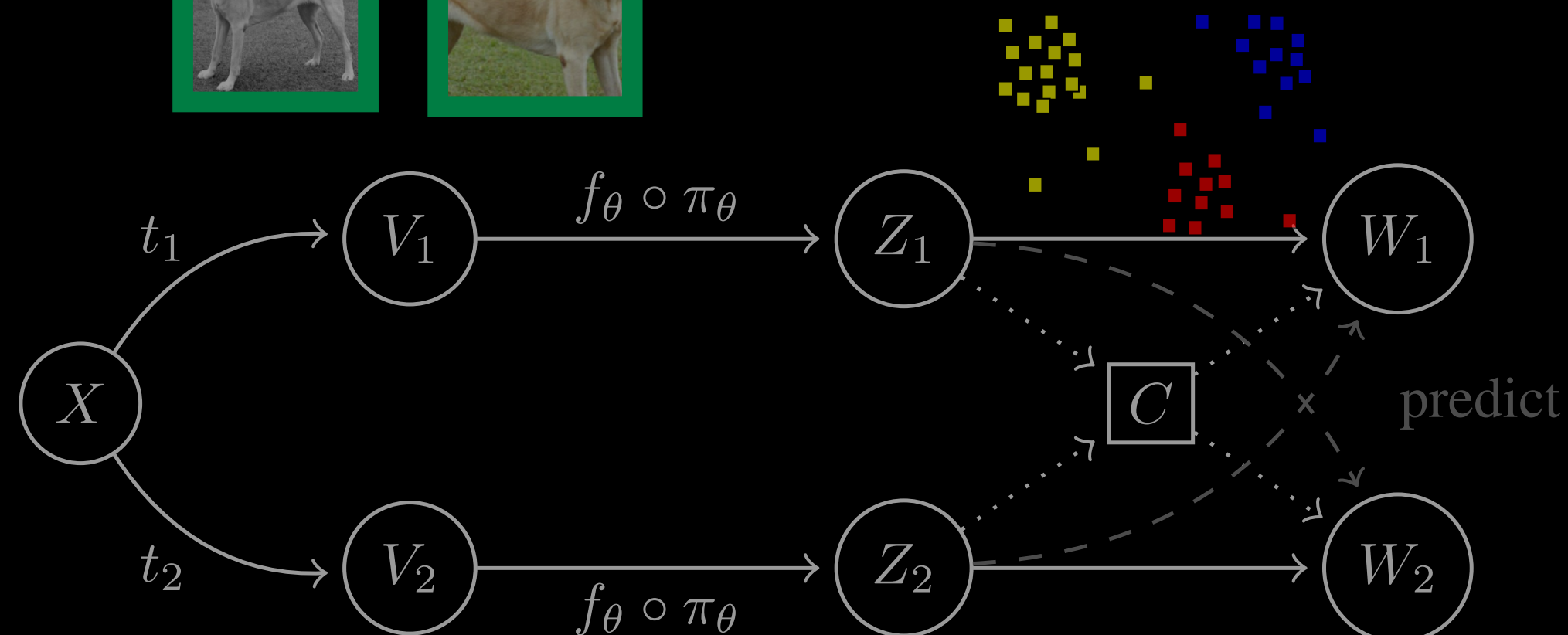
Clustering-based methods

SwAV, DeepCluster

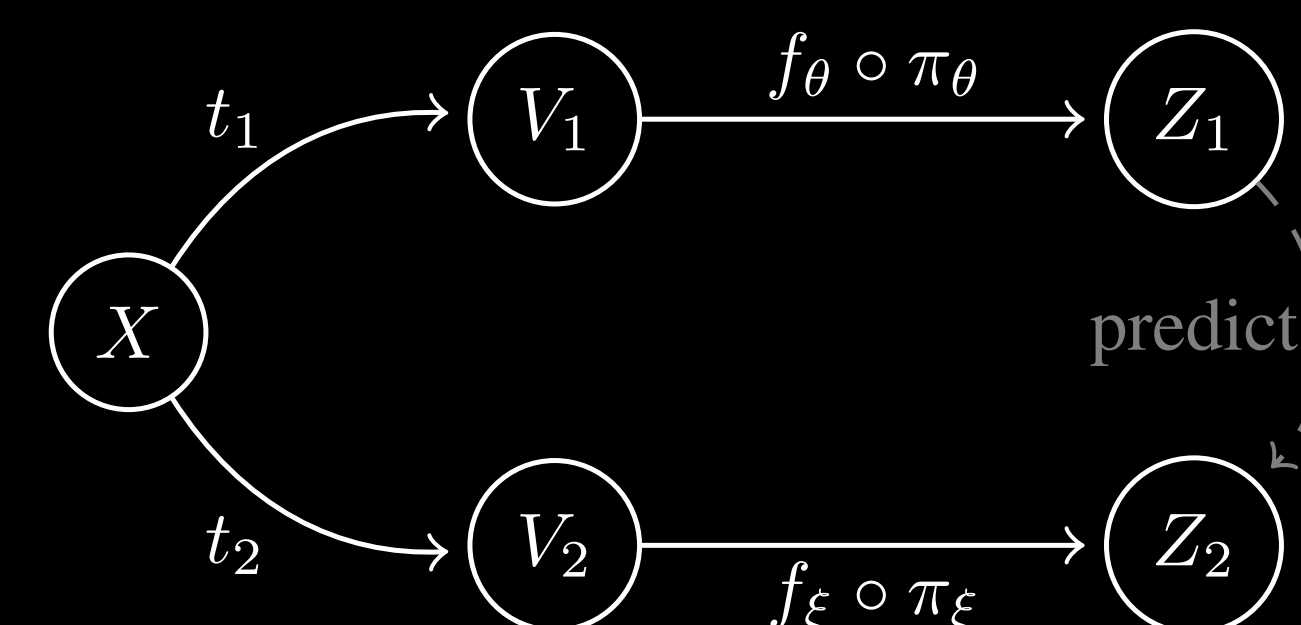
Background: MVSSL families



Contrastive methods
SimCLR, CMC, MoCo

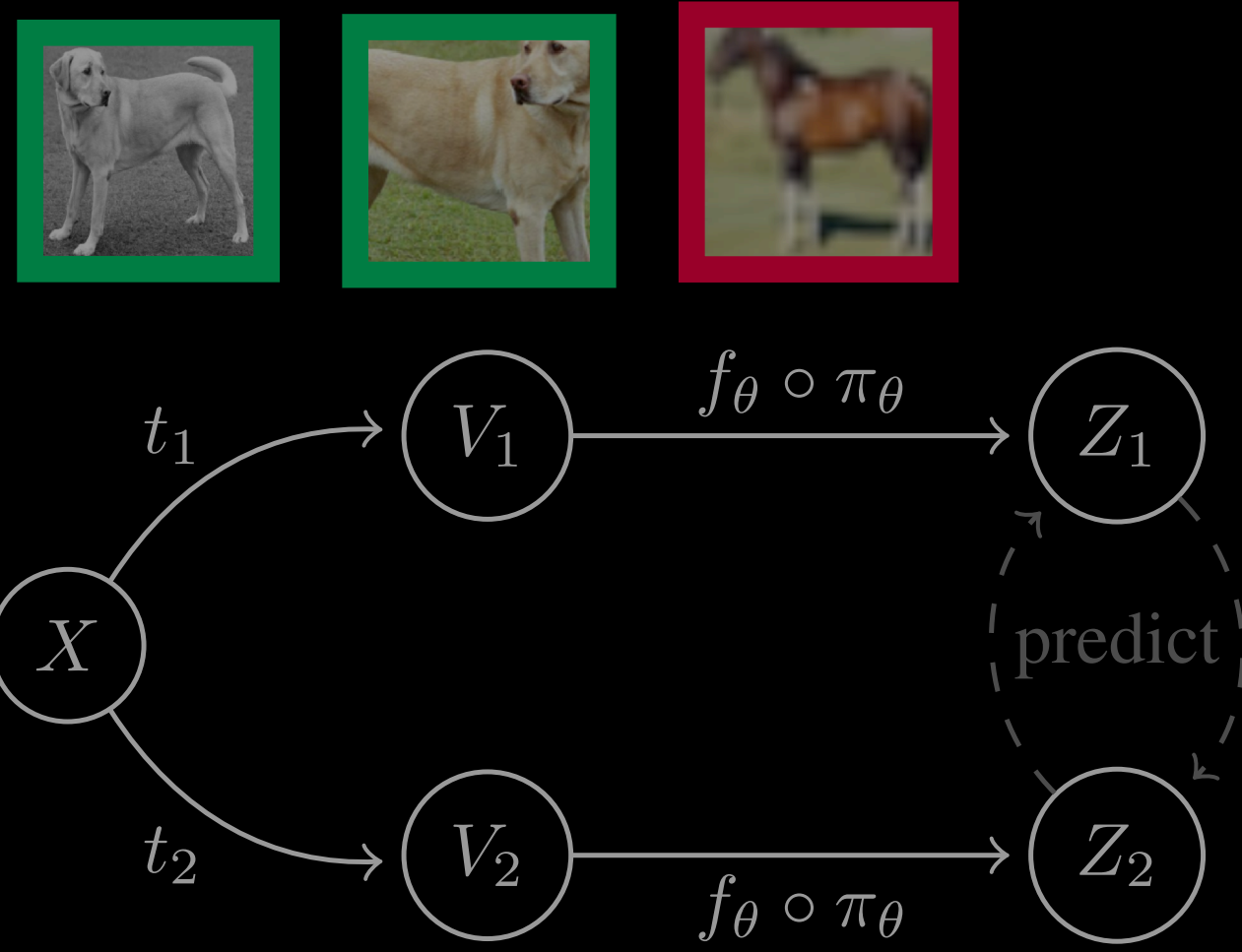


Clustering-based methods
SwAV, DeepCluster

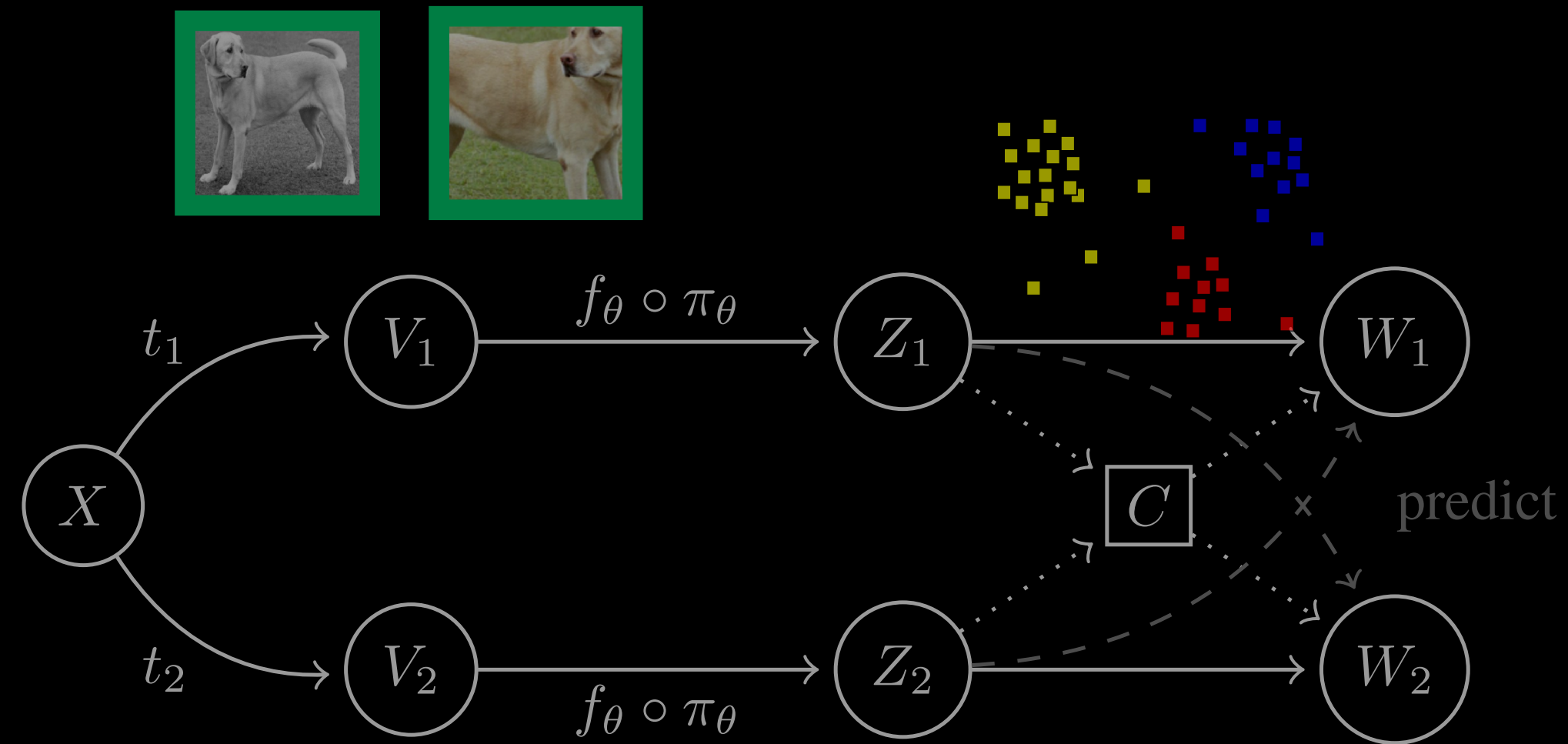


Distillation-based methods
BYOL, DINO

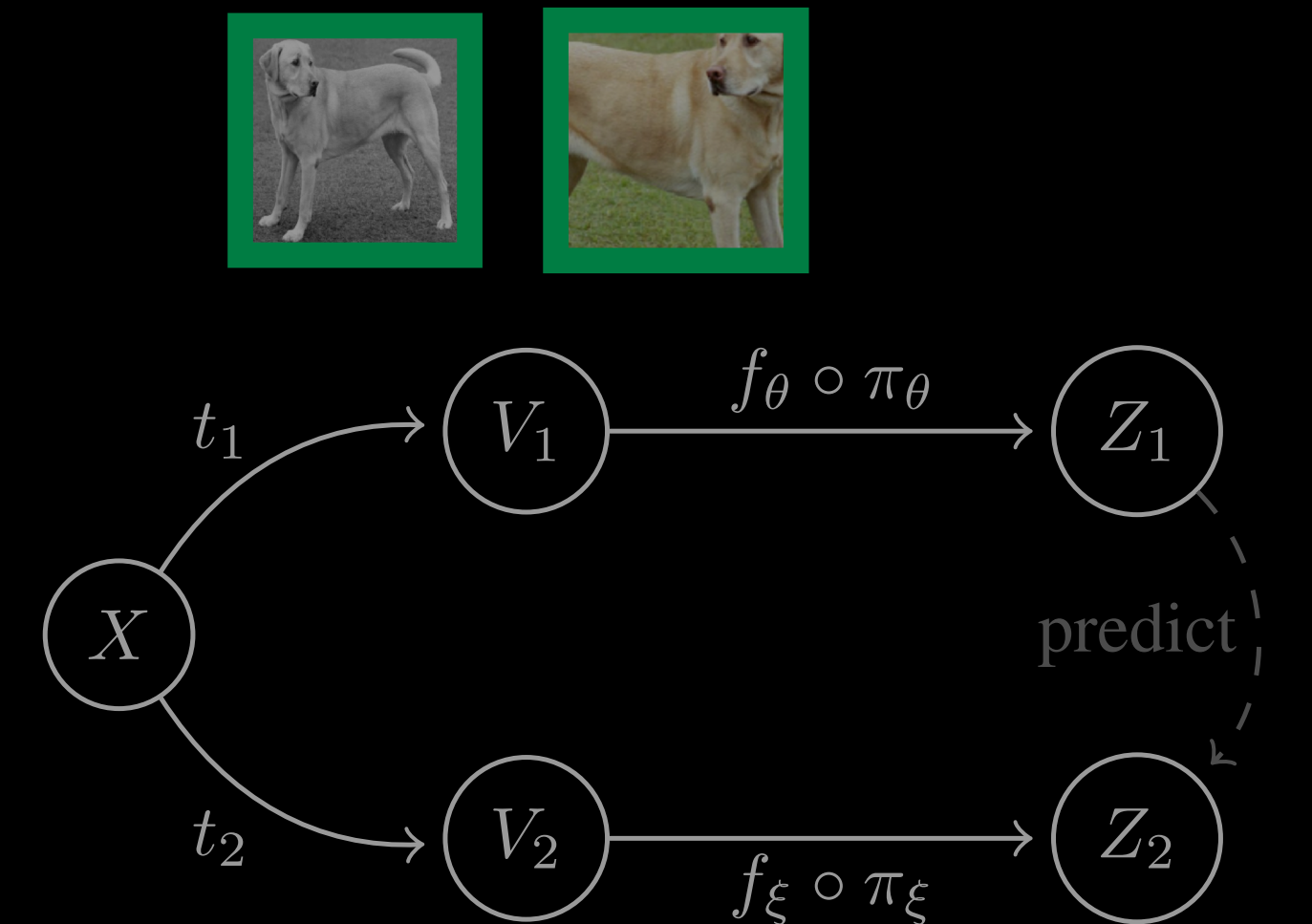
Prior work: What role does MI optimization play in MVSSL?



Contrastive methods
SimCLR, CMC, MoCo

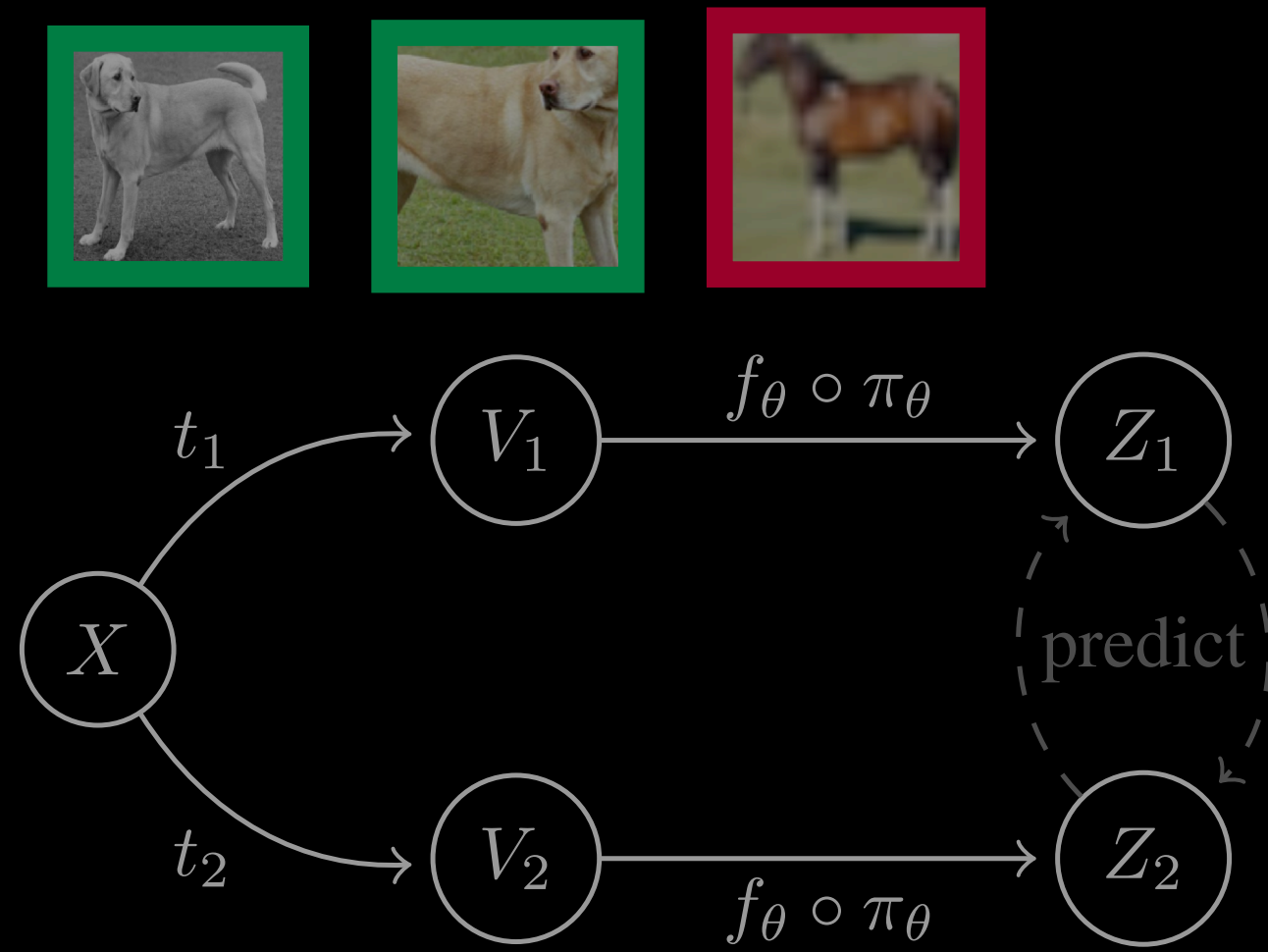


Clustering-based methods
SwAV, DeepCluster



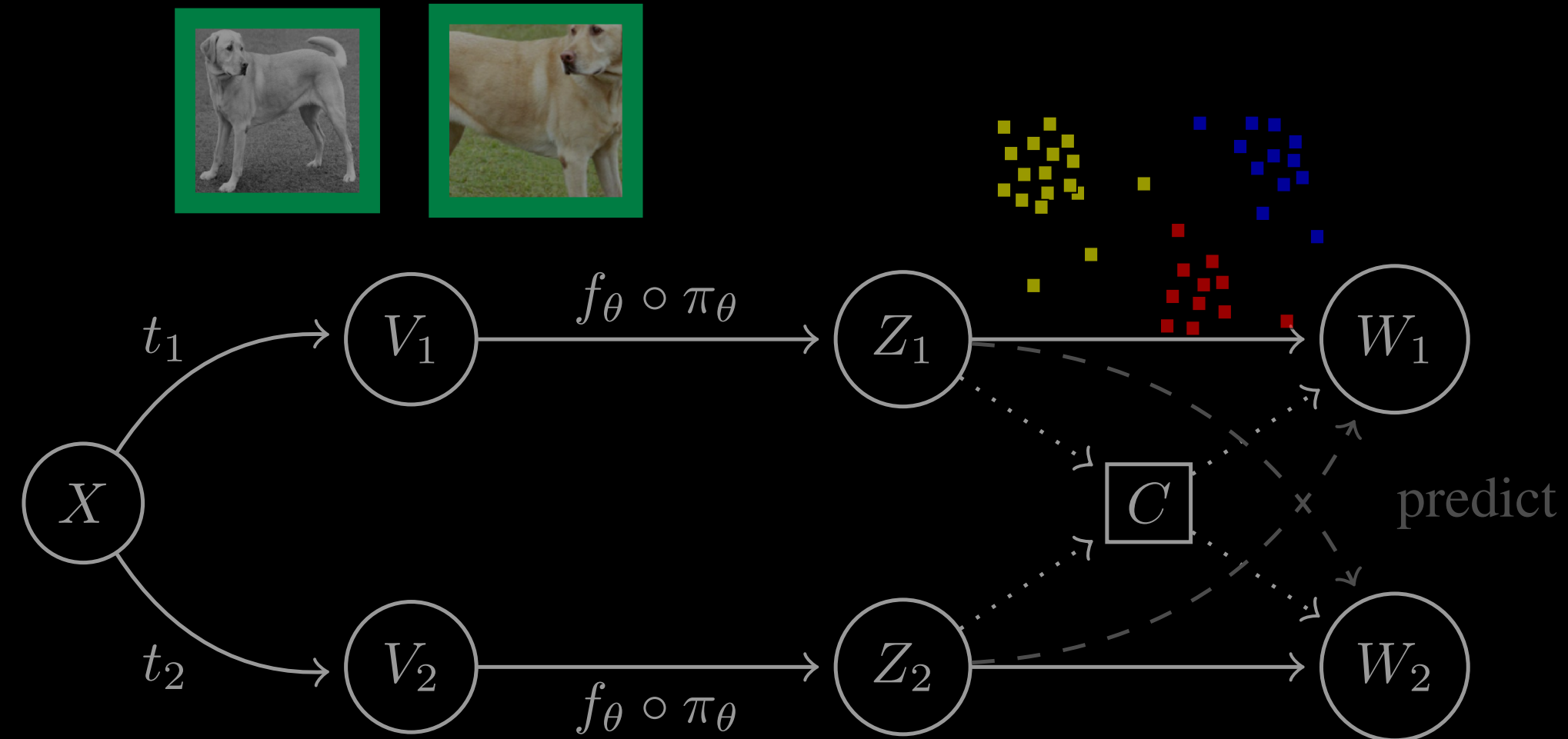
Distillation-based methods
BYOL, DINO

Prior work: What role does MI optimization play in MVSSL?



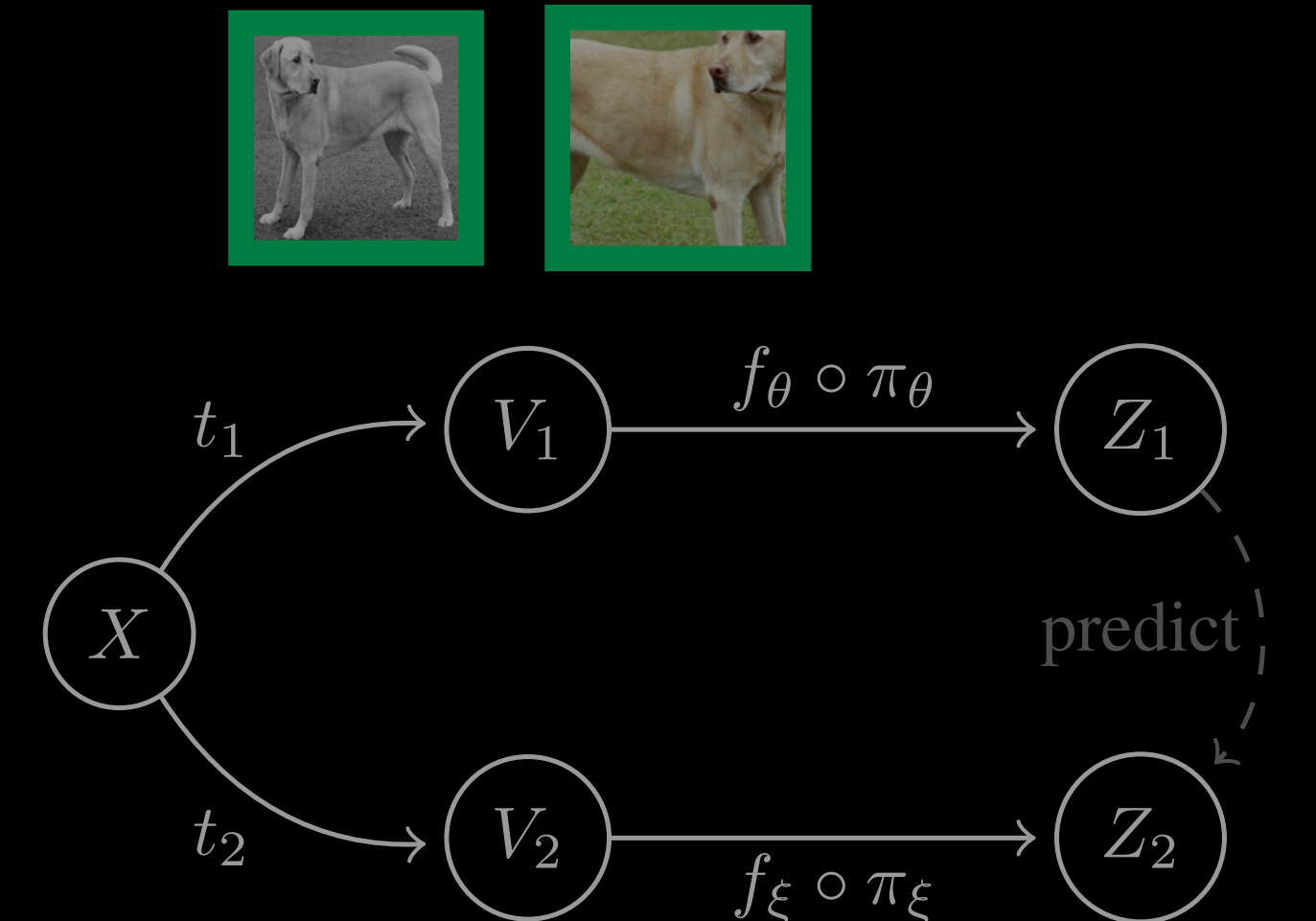
Contrastive methods

SimCLR, CMC, MoCo



Clustering-based methods

SwAV, DeepCluster

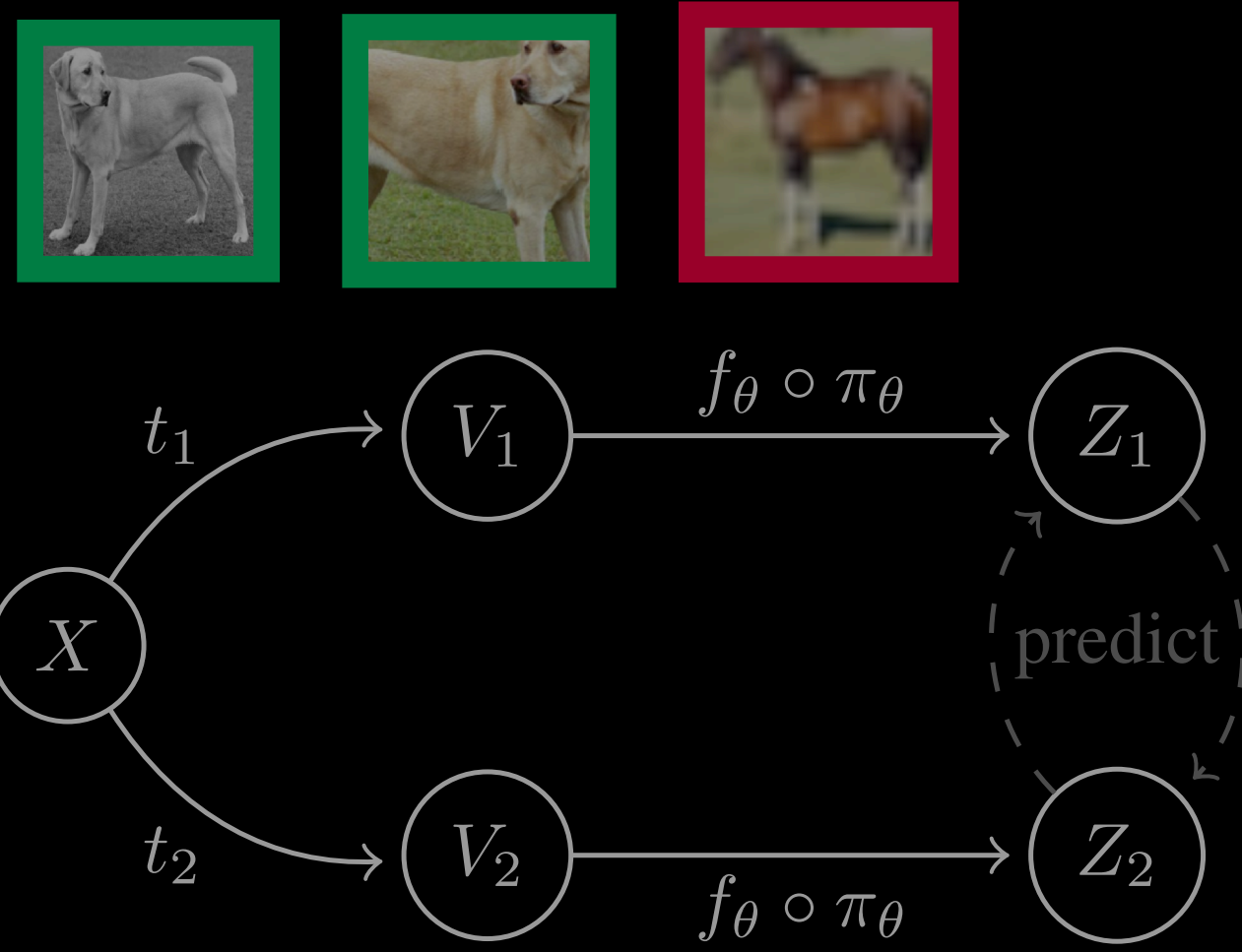


Distillation-based methods

BYOL, DINO

Some optimize $I(Z_1; Z_2)$

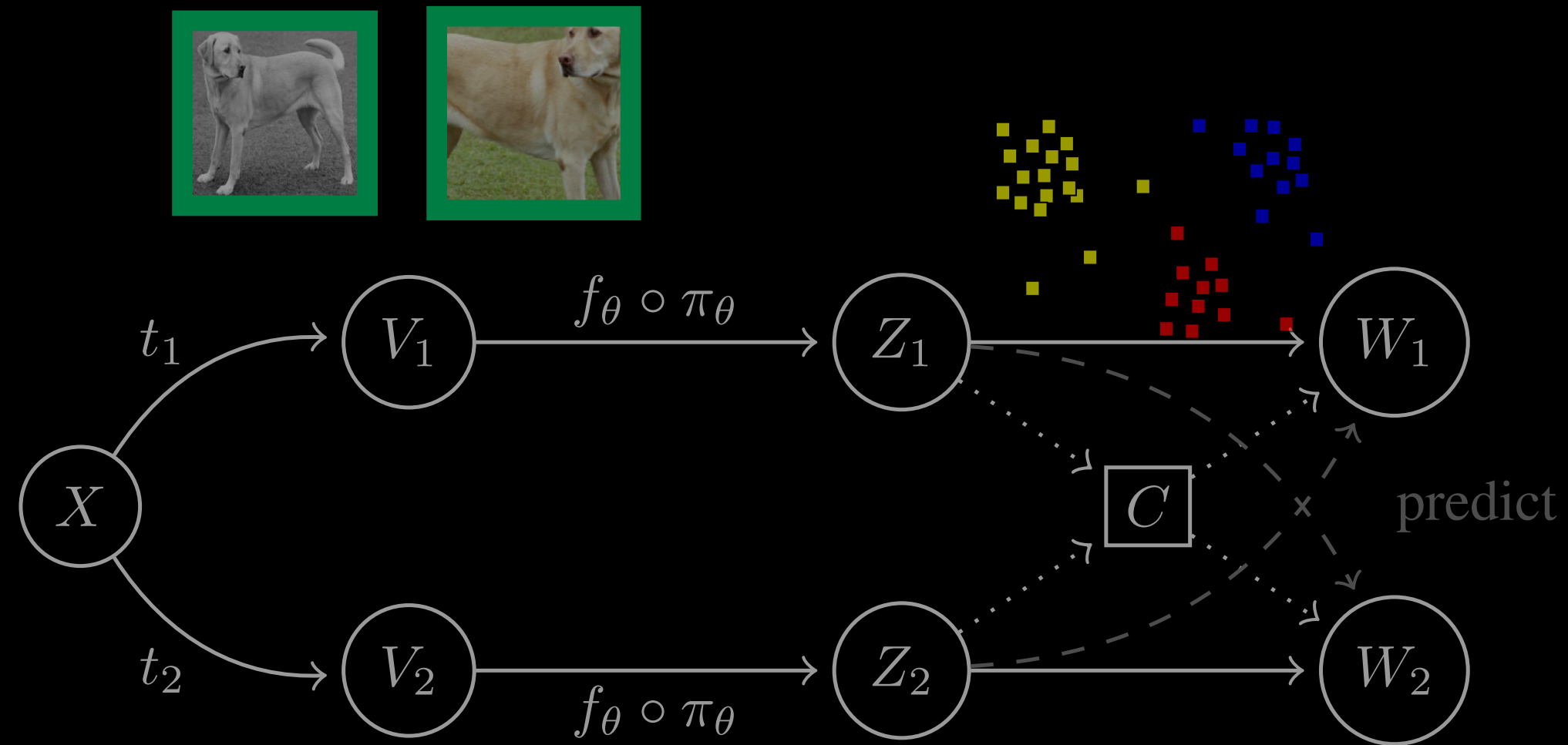
Prior work: What role does MI optimization play in MVSSL?



Contrastive methods

SimCLR, CMC, MoCo

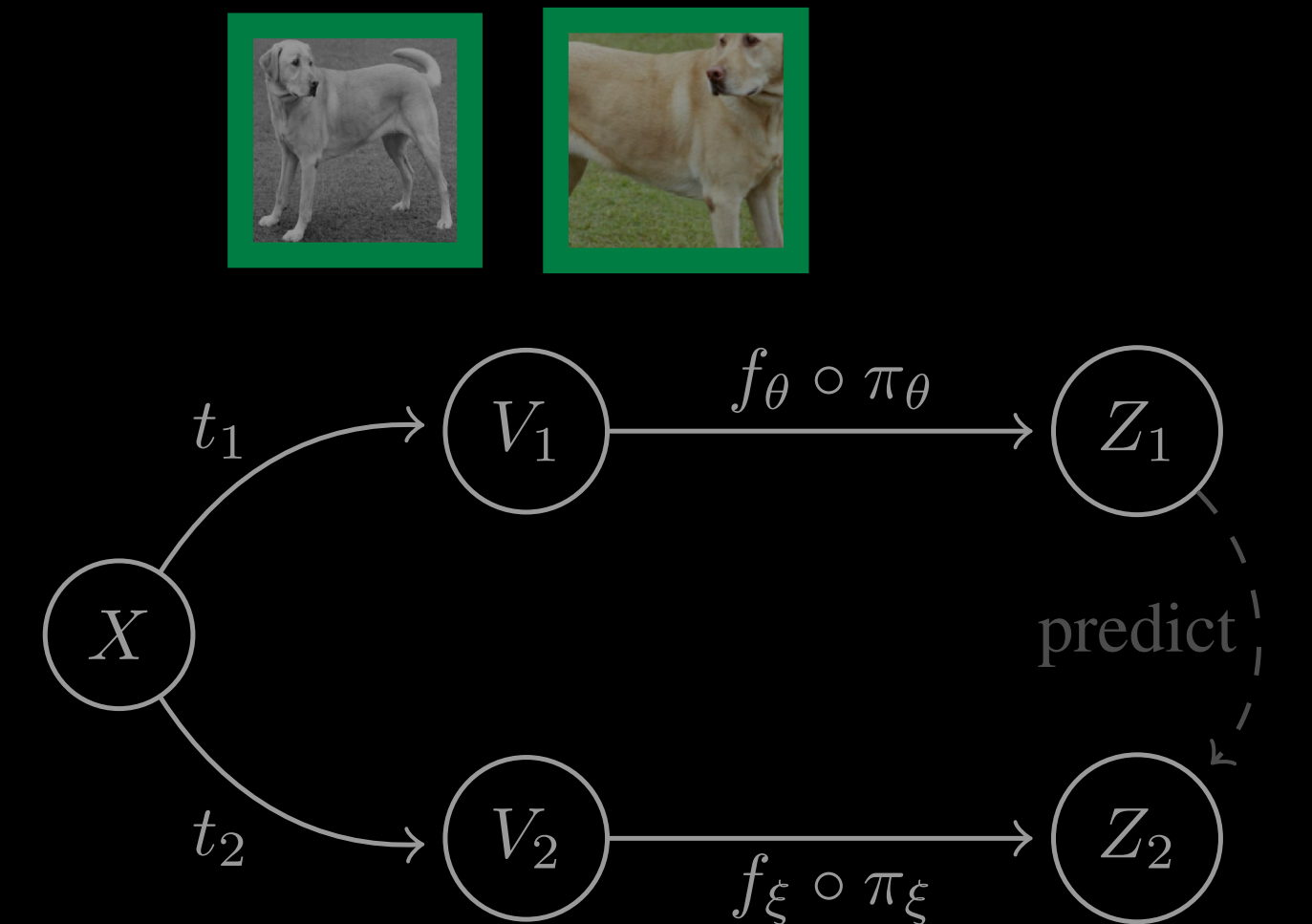
Some optimize $I(Z_1; Z_2)$



Clustering-based methods

SwAV, DeepCluster

?



Distillation-based methods

BYOL, DINO

?

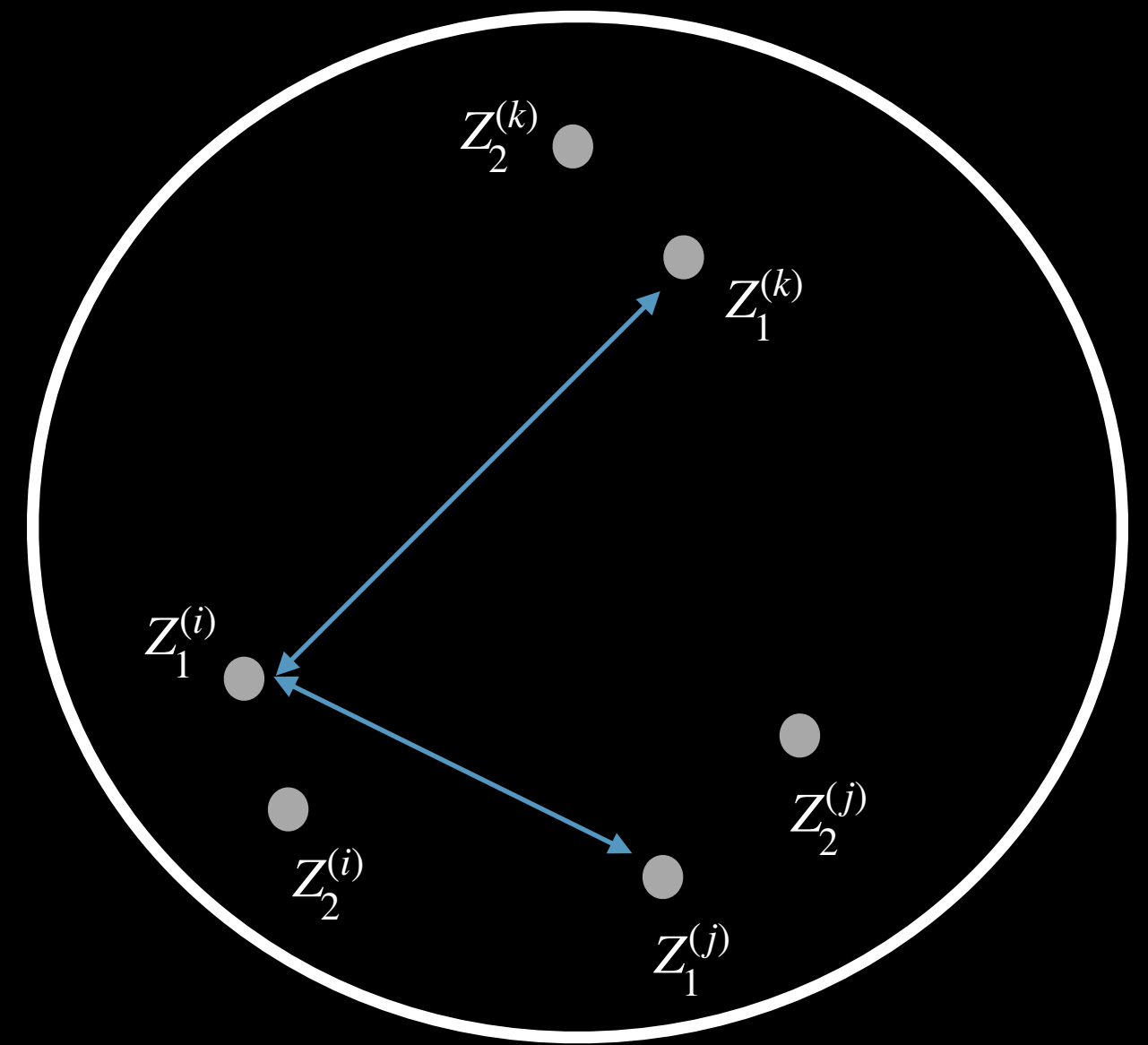
Analysis: Using a different bound on MI w.r.t. prior works

$$\begin{aligned} I(Z_1; Z_2) &= H(Z_2) - H(Z_2 | Z_1) \\ &\geq H(Z_2) - \mathbb{E}_{Z_1, Z_2} \left[-\log q_{Z_2|Z_1}(Z_2) \right] := I_{\text{ER}}(Z_1; Z_2) \end{aligned}$$

Analysis: Using a different bound on MI w.r.t. prior works

$$\begin{aligned} I(Z_1; Z_2) &= H(Z_2) - H(Z_2 | Z_1) \\ &\geq \boxed{H(Z_2)} - \mathbb{E}_{Z_1, Z_2} \left[-\log q_{Z_2|Z_1}(Z_2) \right] := I_{\text{ER}}(Z_1; Z_2) \end{aligned}$$

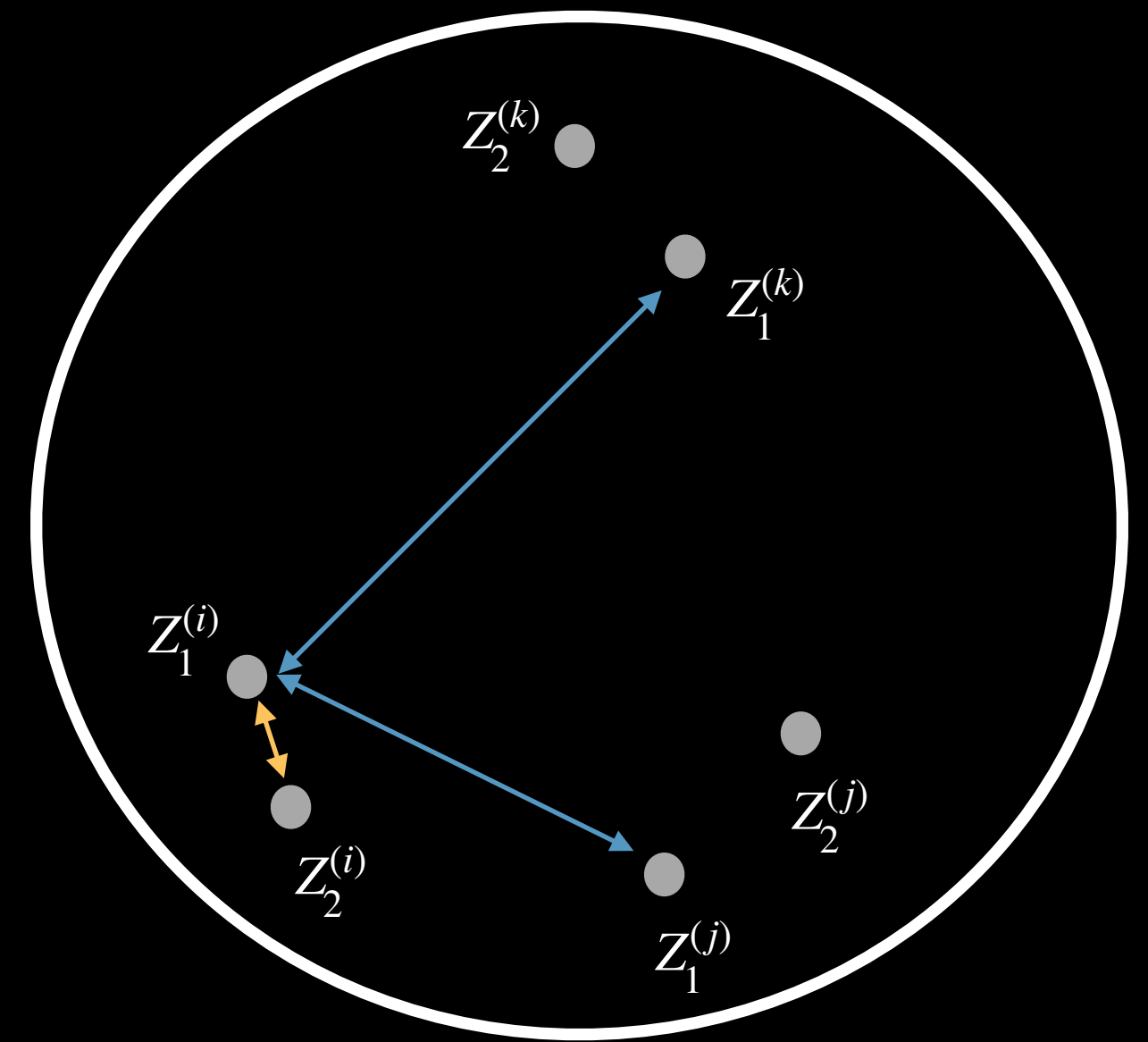
- Entropy: How much information *can* be learnt



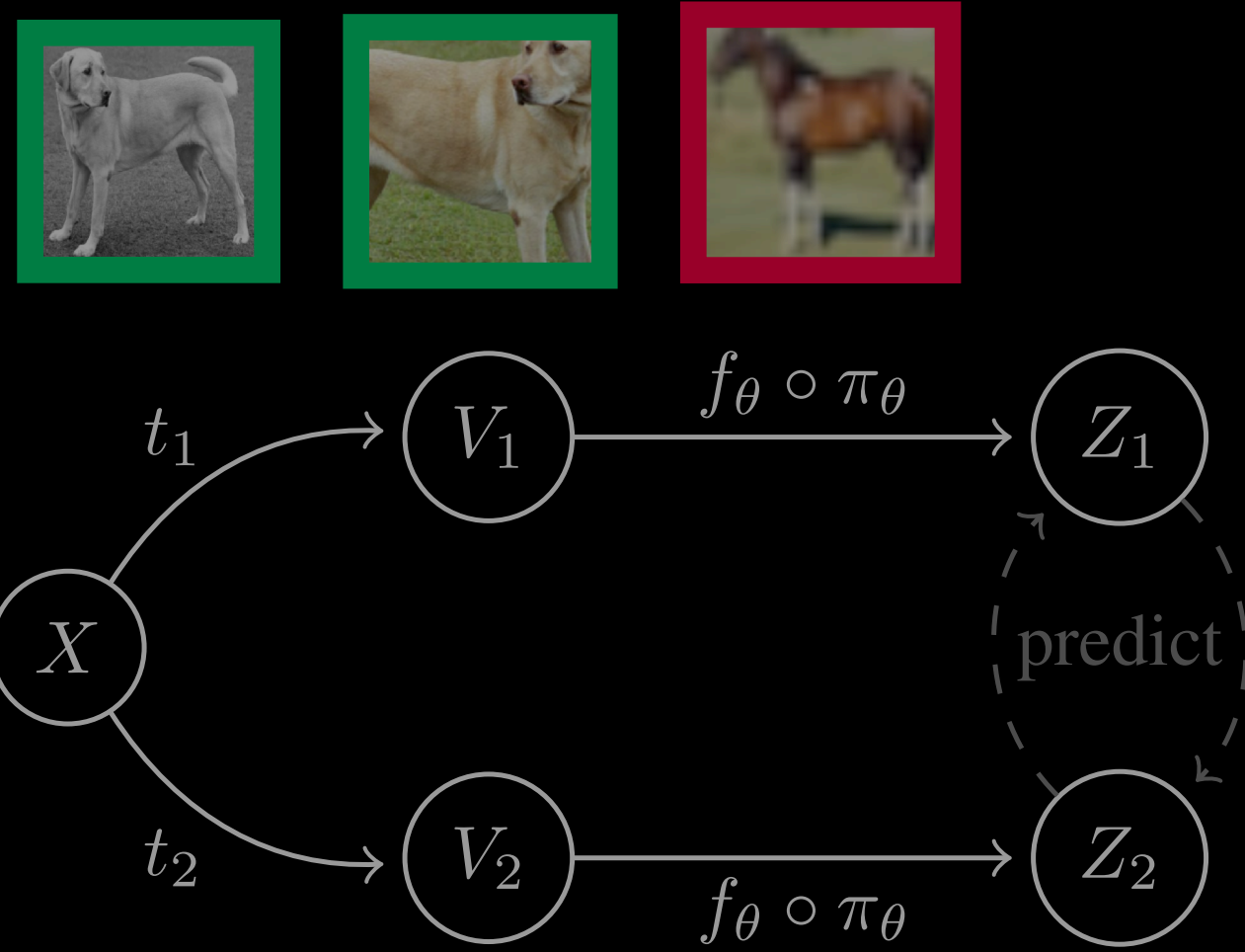
Analysis: Using a different bound on MI w.r.t. prior works

$$\begin{aligned} I(Z_1; Z_2) &= H(Z_2) - H(Z_2 | Z_1) \\ &\geq \boxed{H(Z_2)} - \boxed{\mathbb{E}_{Z_1, Z_2} \left[-\log q_{Z_2|Z_1}(Z_2) \right]} := I_{\text{ER}}(Z_1; Z_2) \end{aligned}$$

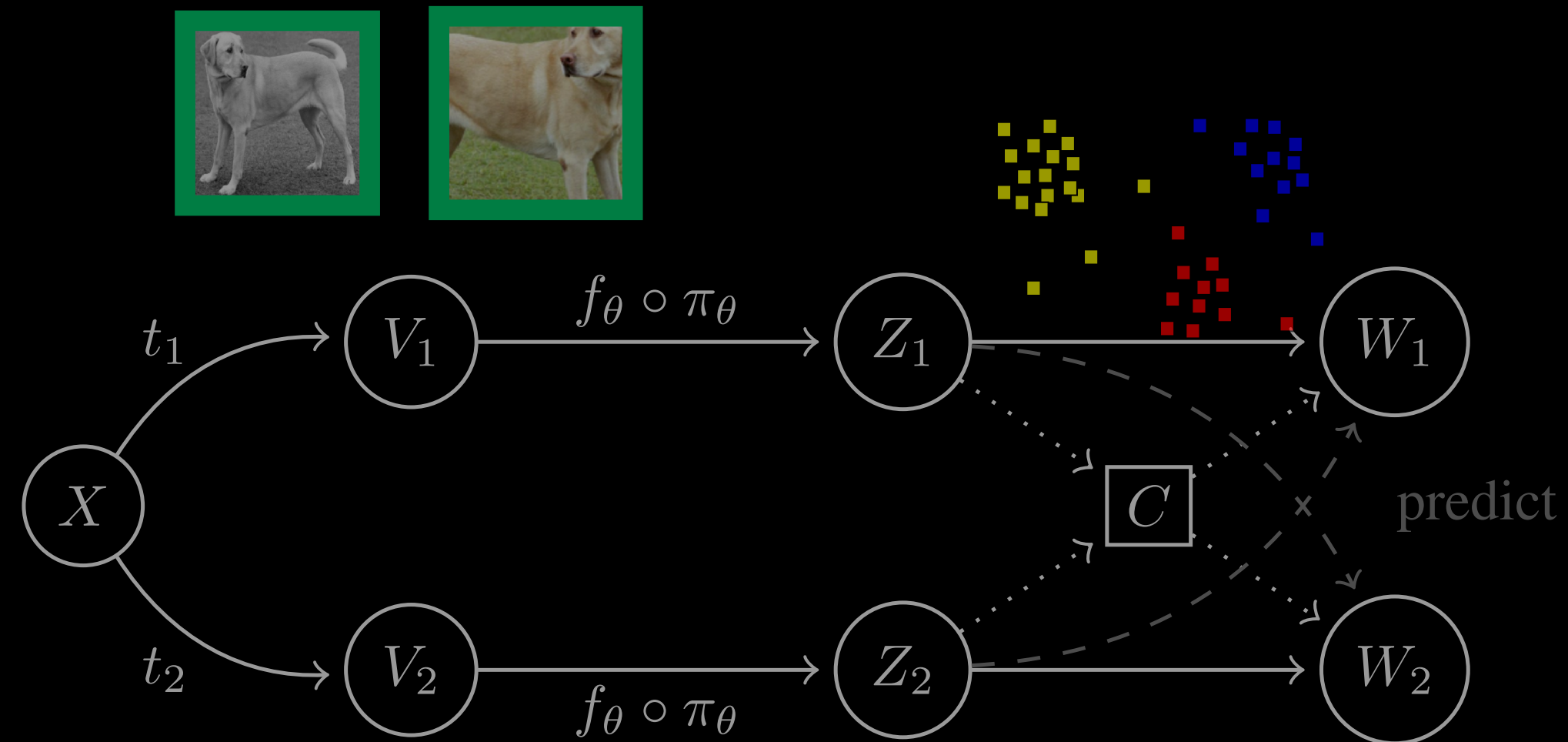
- **Entropy**: How much information *can* be learnt
- **Reconstruction**: How much information *is* learnt



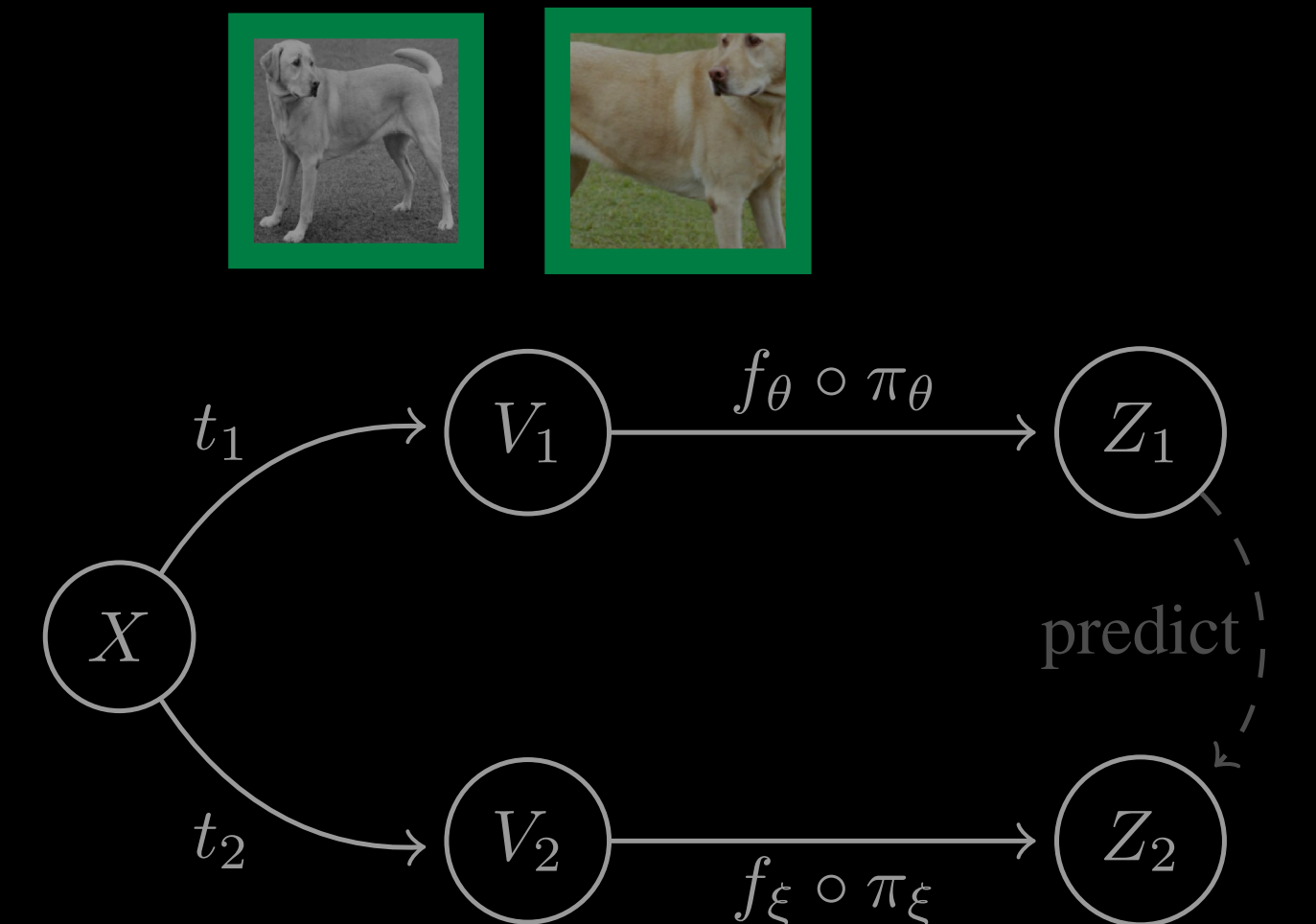
Theoretical analysis



Contrastive methods
SimCLR, CMC, MoCo

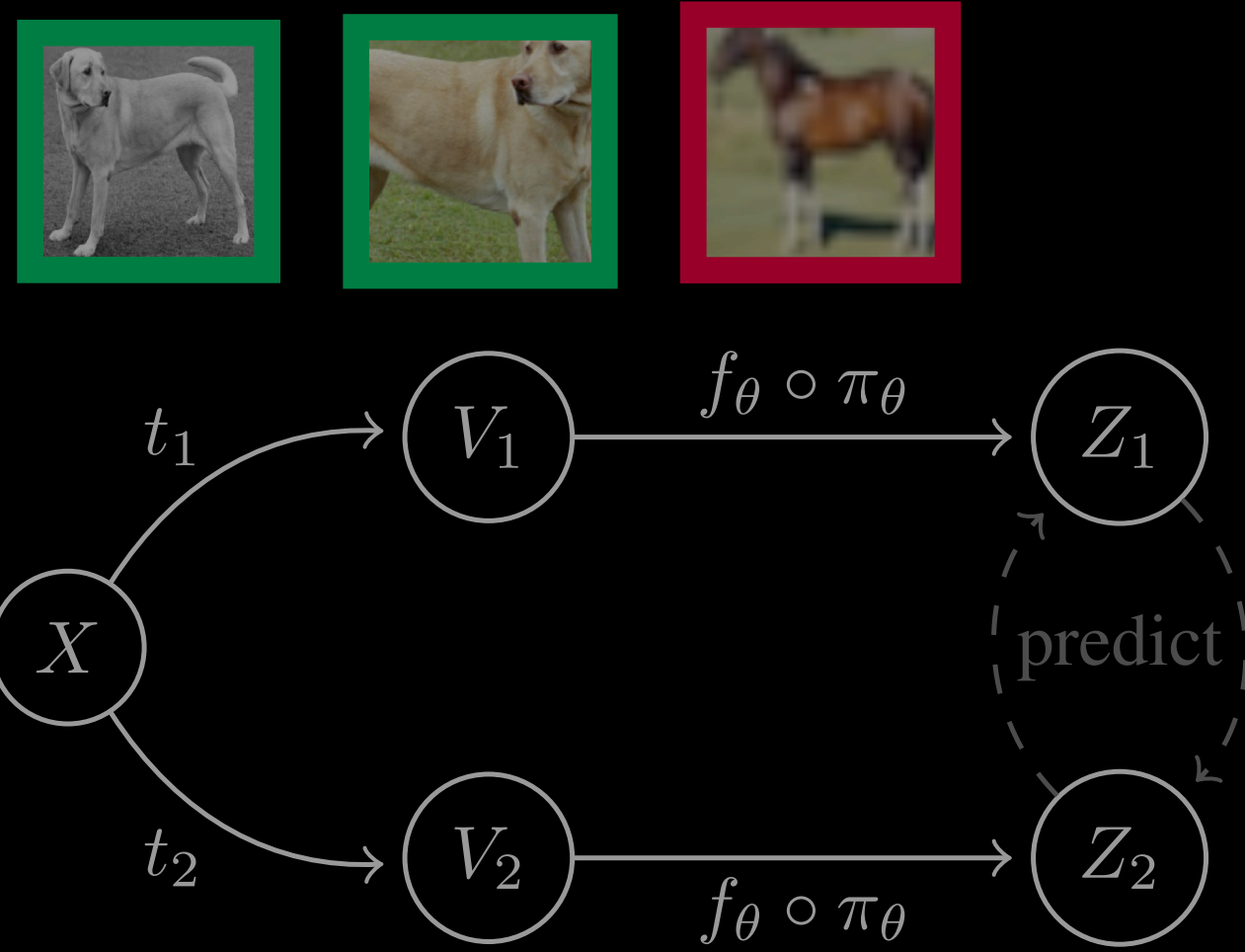


Clustering-based methods
SwAV, DeepCluster



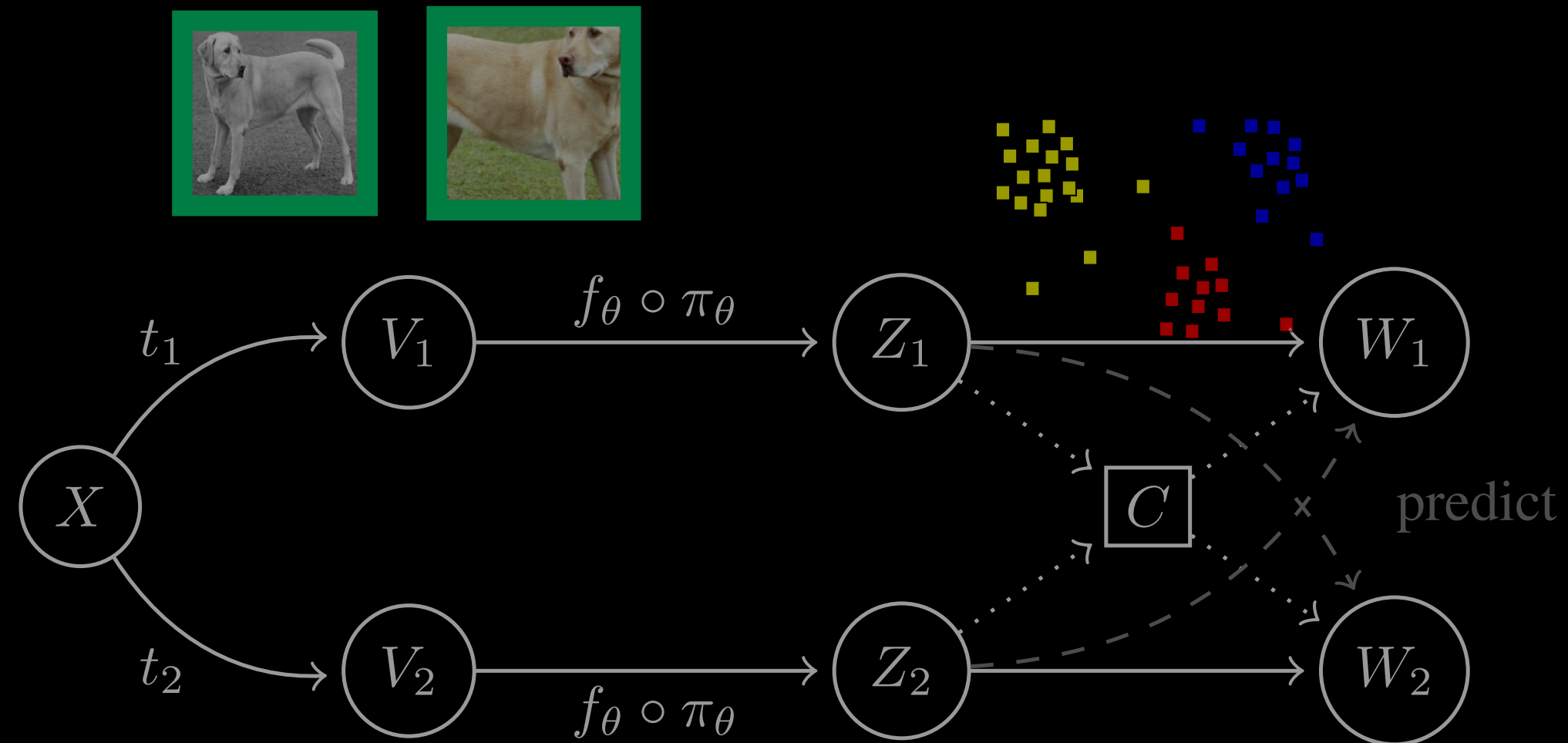
Distillation-based methods
BYOL, DINO

Theoretical analysis



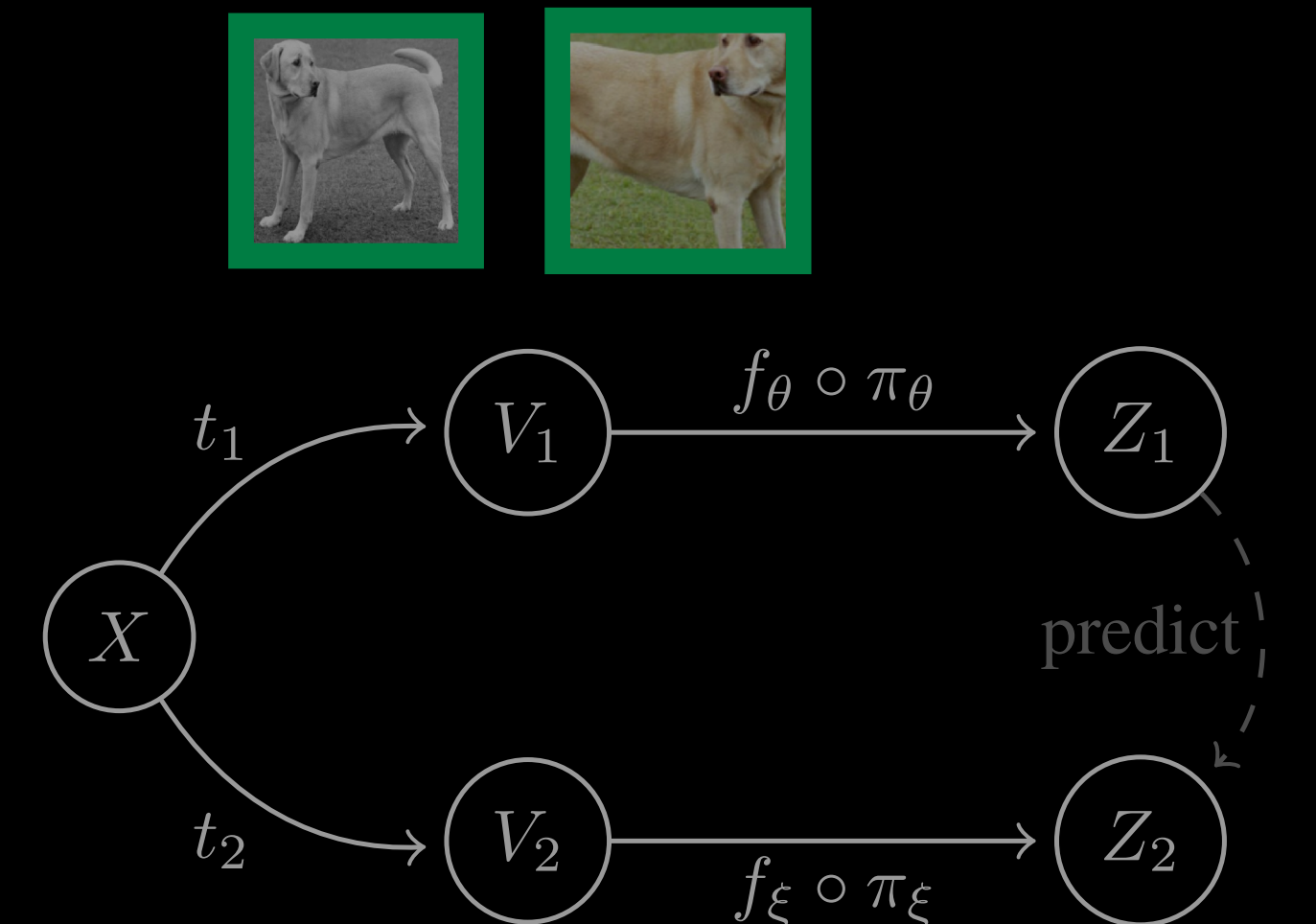
Contrastive methods

SimCLR, CMC, MoCo



Clustering-based methods

SwAV, DeepCluster

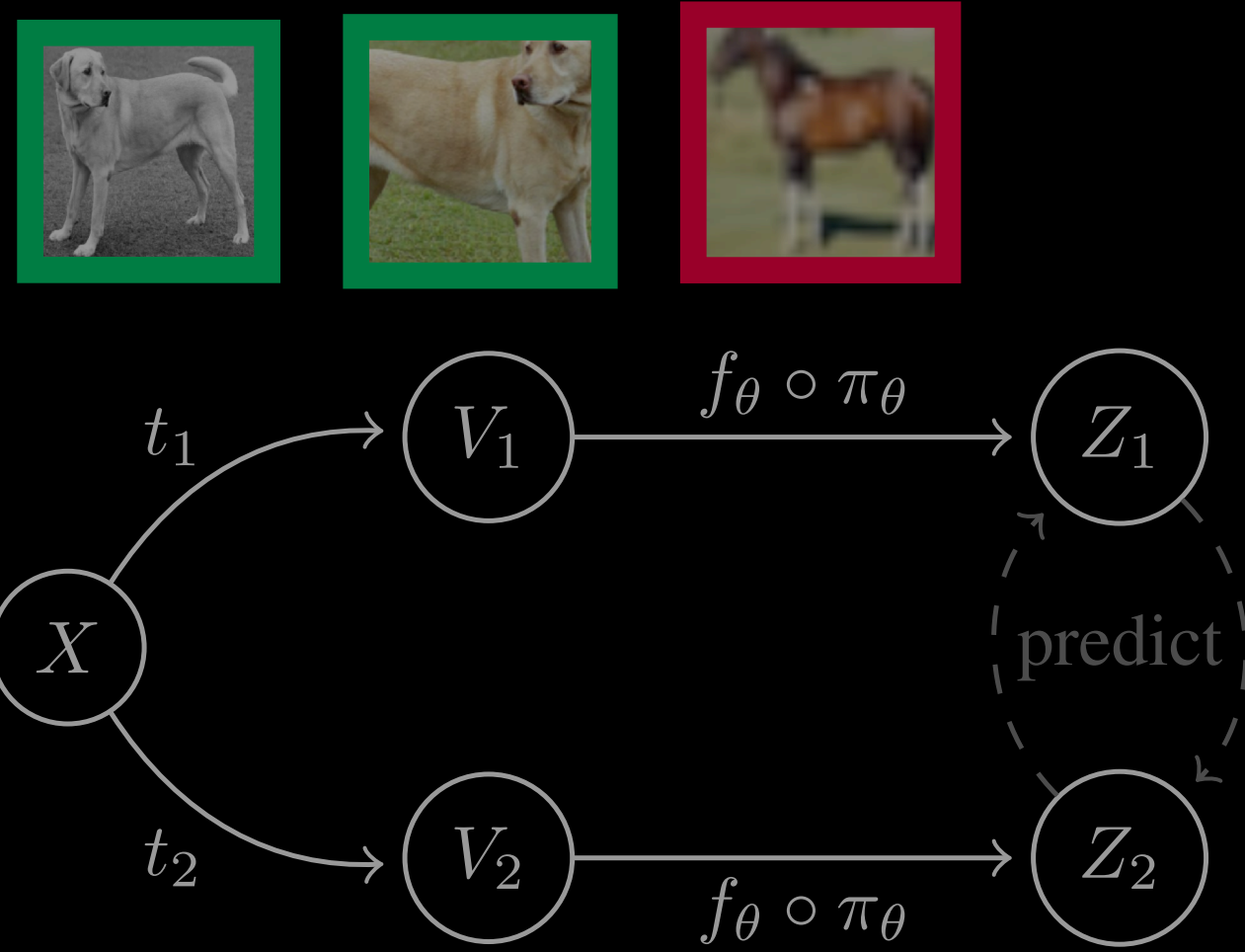


Distillation-based methods

BYOL, DINO

Some optimize $I(Z_1; Z_2)$
exactly, some not exactly

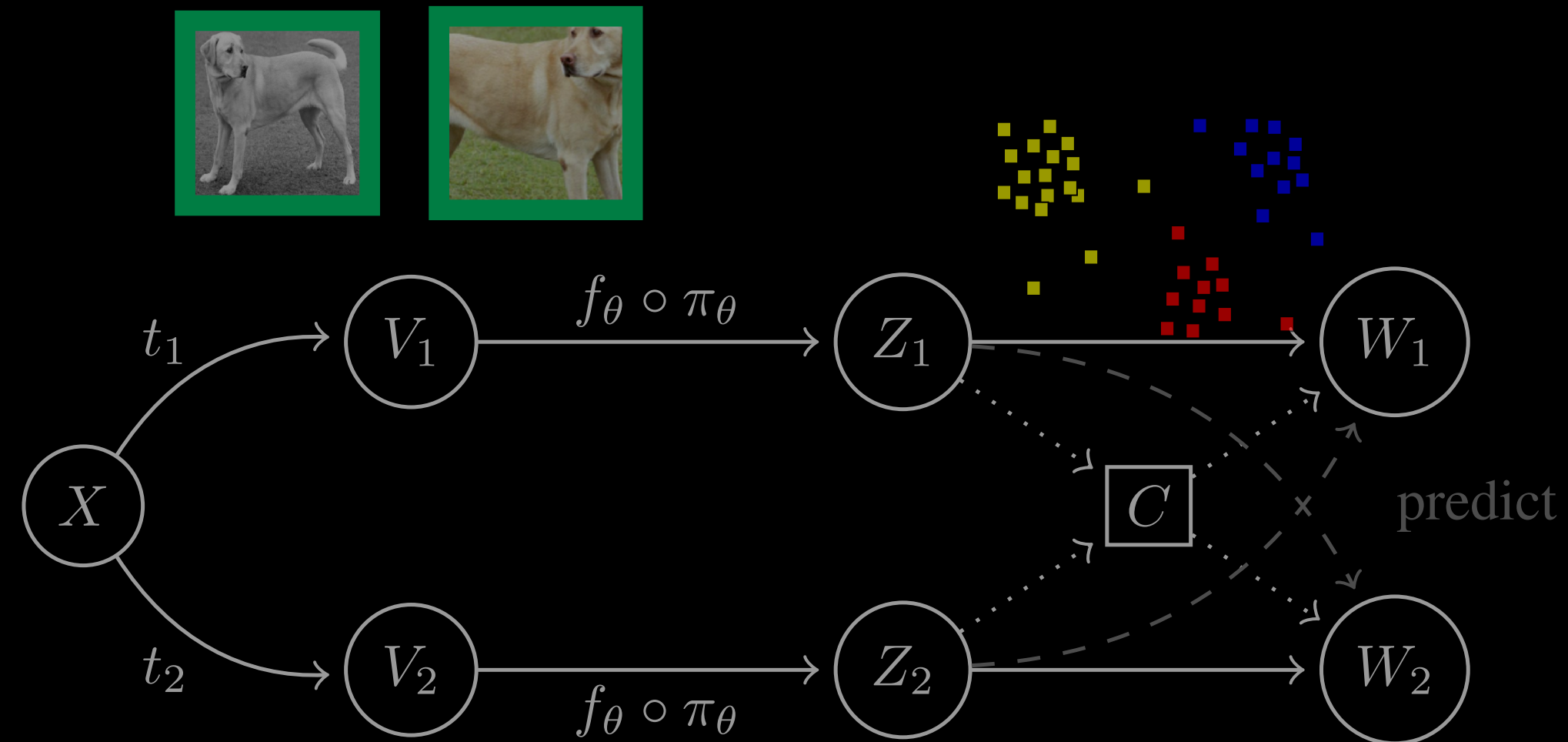
Theoretical analysis



Contrastive methods

SimCLR, CMC, MoCo

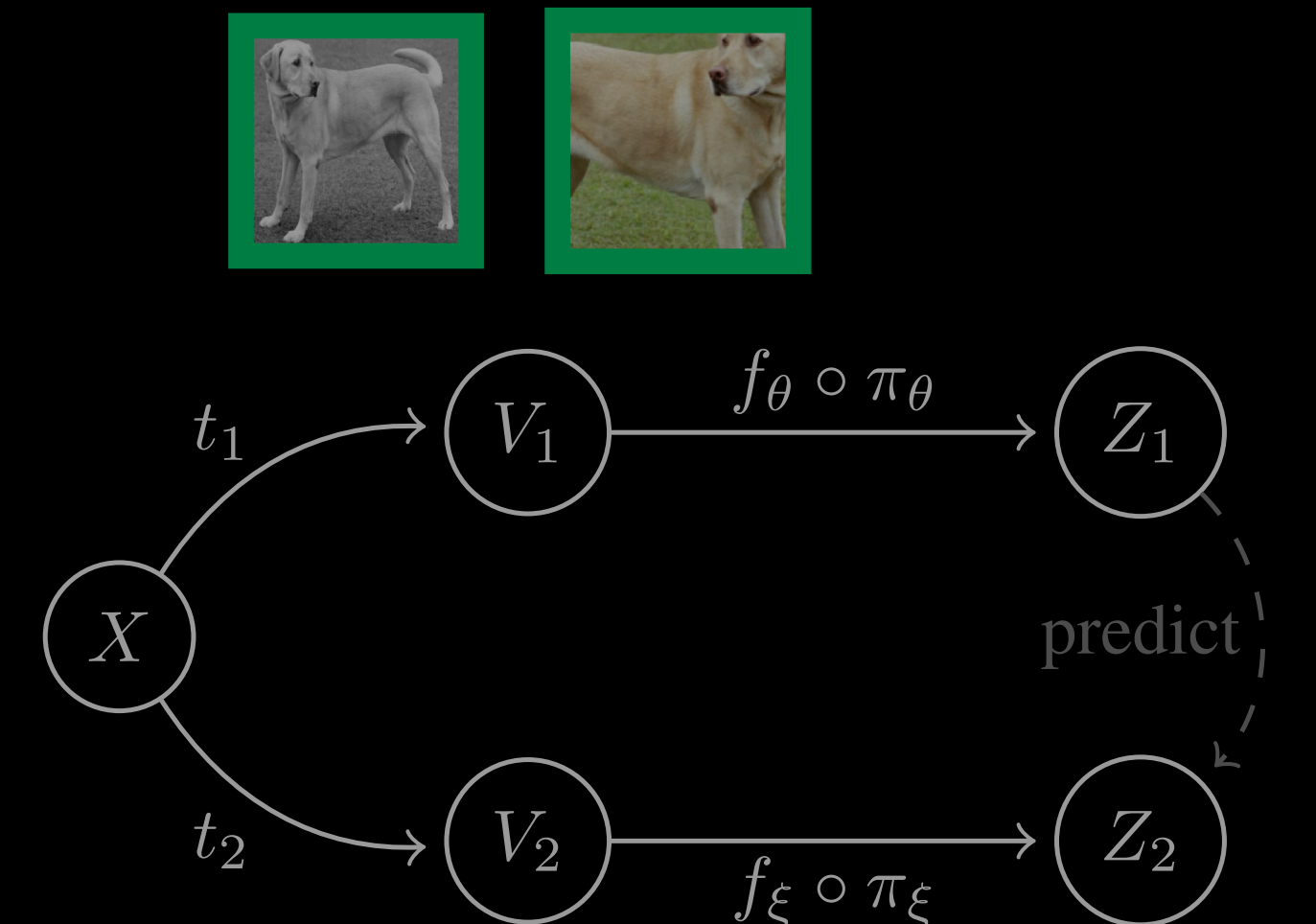
Some optimize $I(Z_1; Z_2)$ exactly, some not exactly



Clustering-based methods

SwAV, DeepCluster

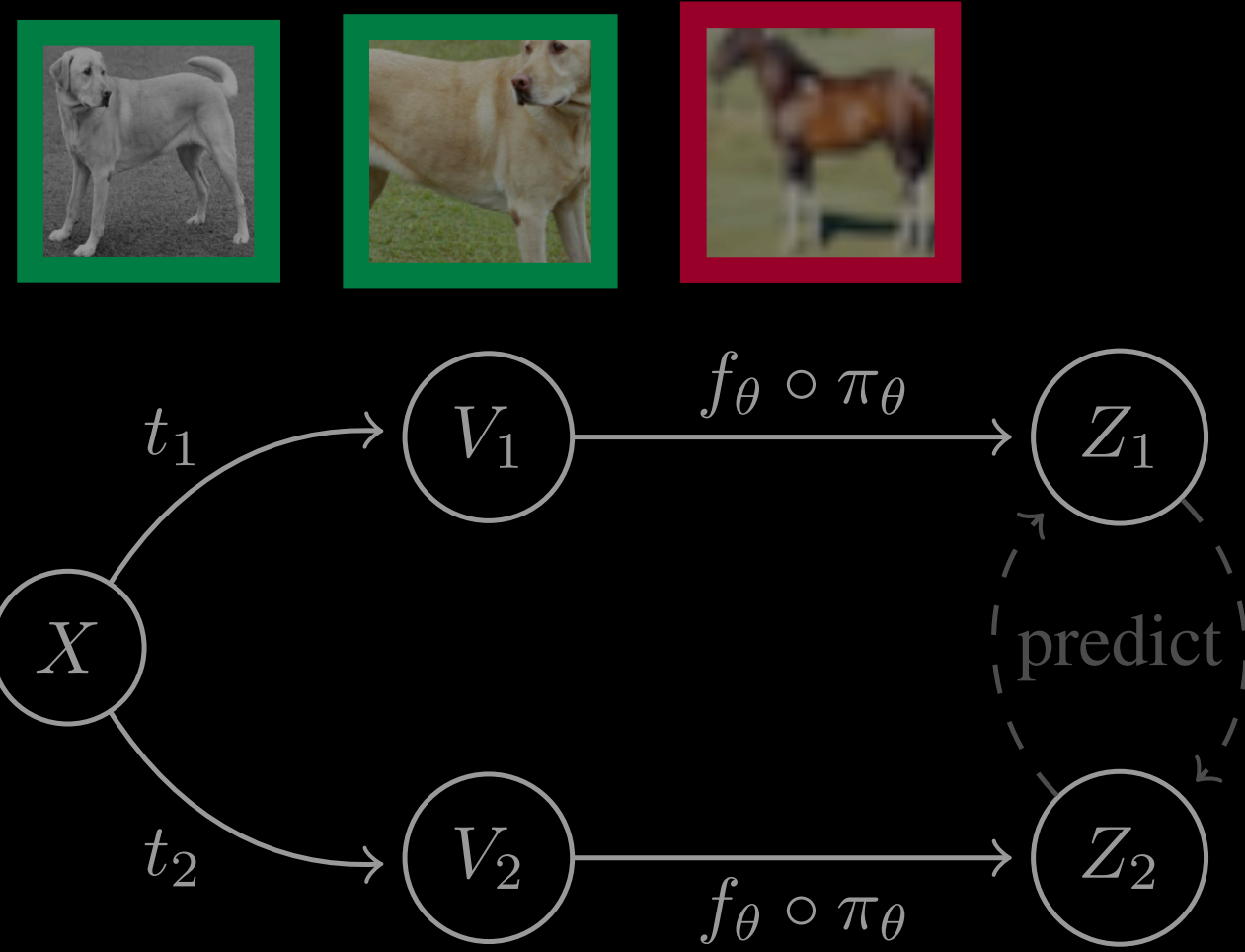
Optimize $I(Z_1; Z_2)$ exactly



Distillation-based methods

BYOL, DINO

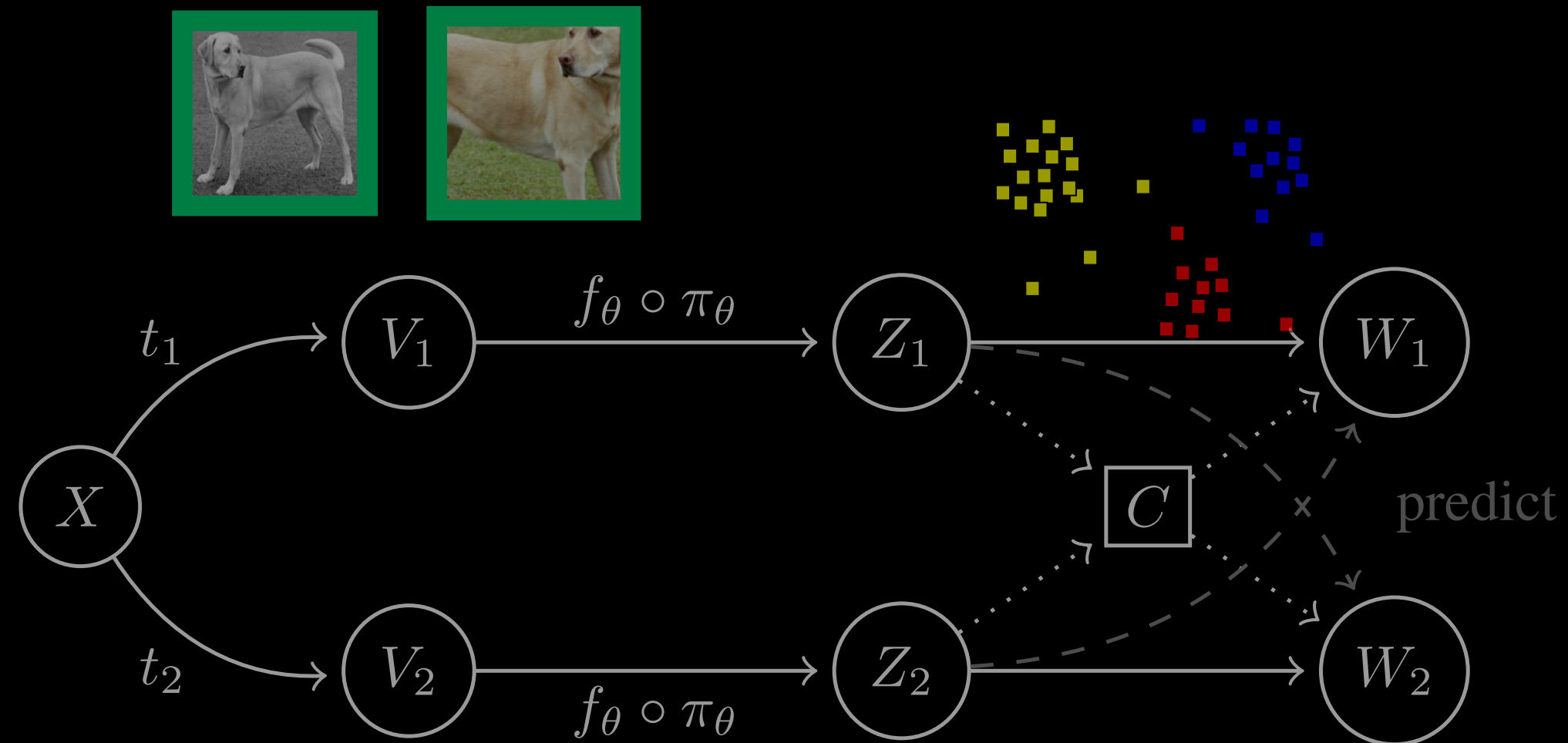
Theoretical analysis



Contrastive methods

SimCLR, CMC, MoCo

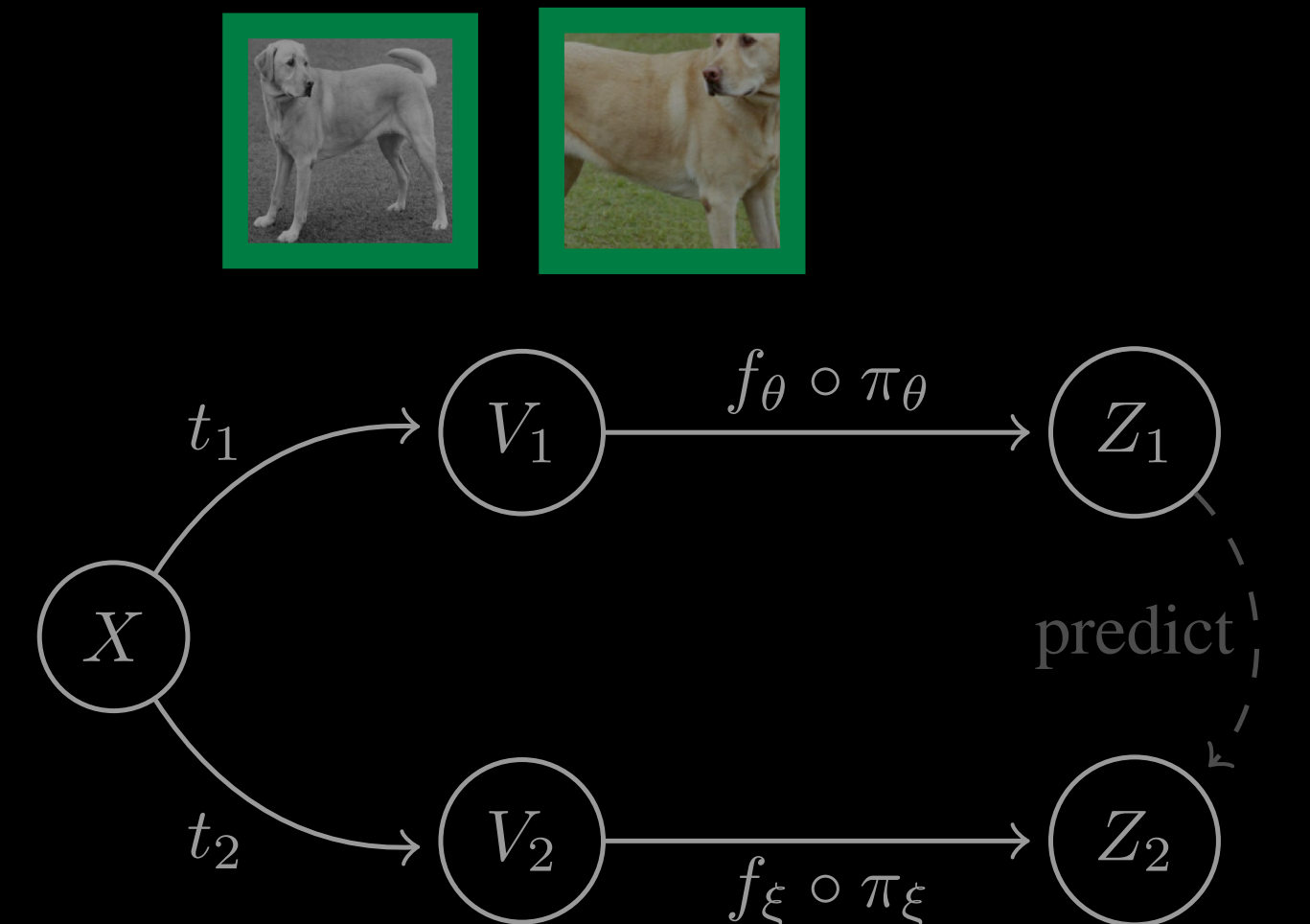
Some optimize $I(Z_1; Z_2)$ exactly, some not exactly



Clustering-based methods

SwAV, DeepCluster

Optimize $I(Z_1; Z_2)$ exactly



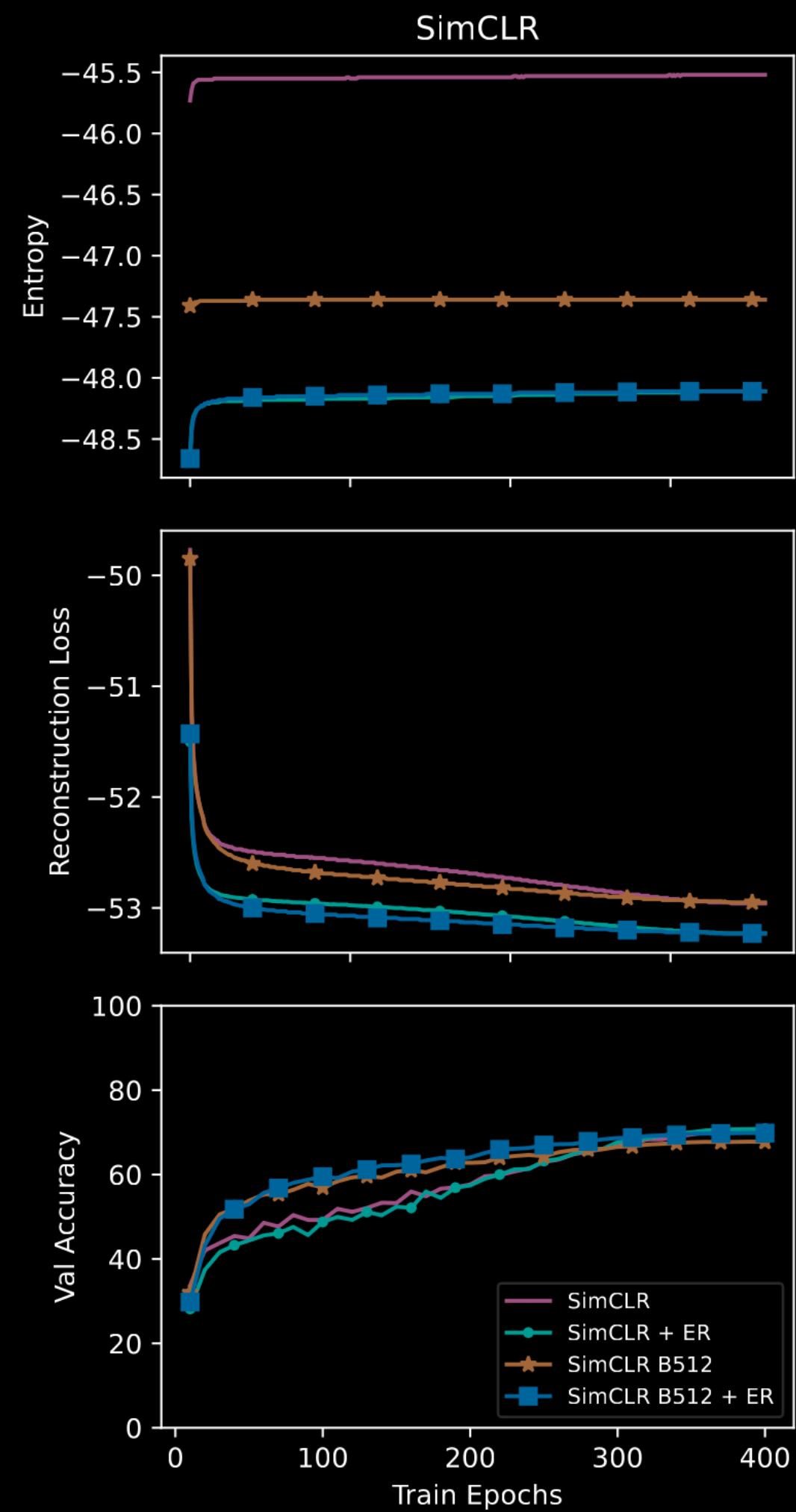
Distillation-based methods

BYOL, DINO

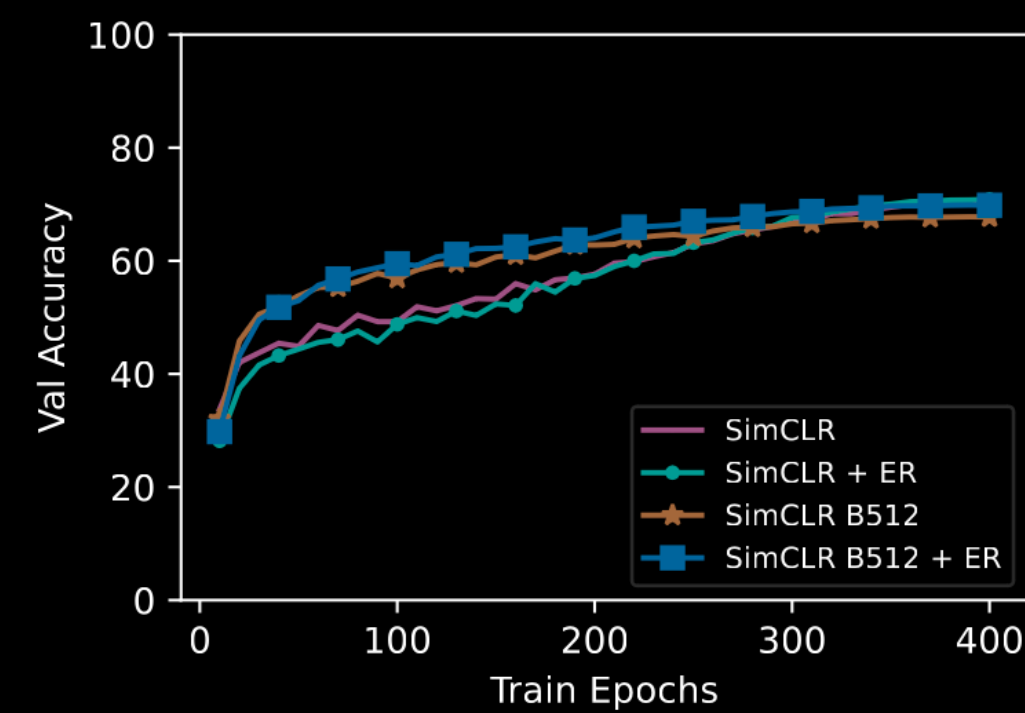
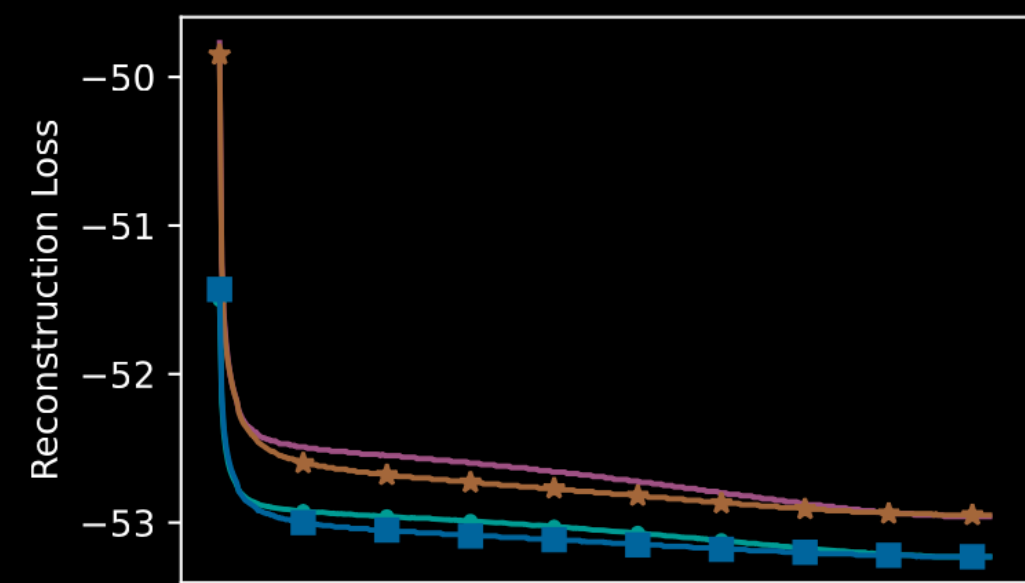
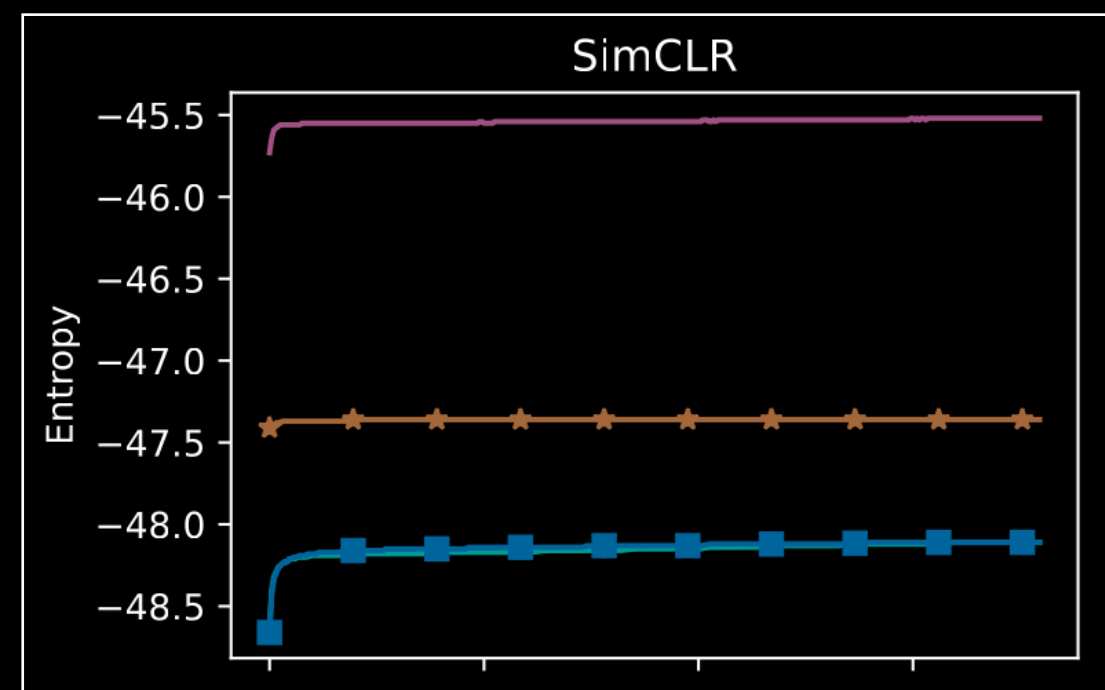
Maximize reconstruction, but the entropy is only maintained stable

Empirical results: We can add entropy optimization explicitly

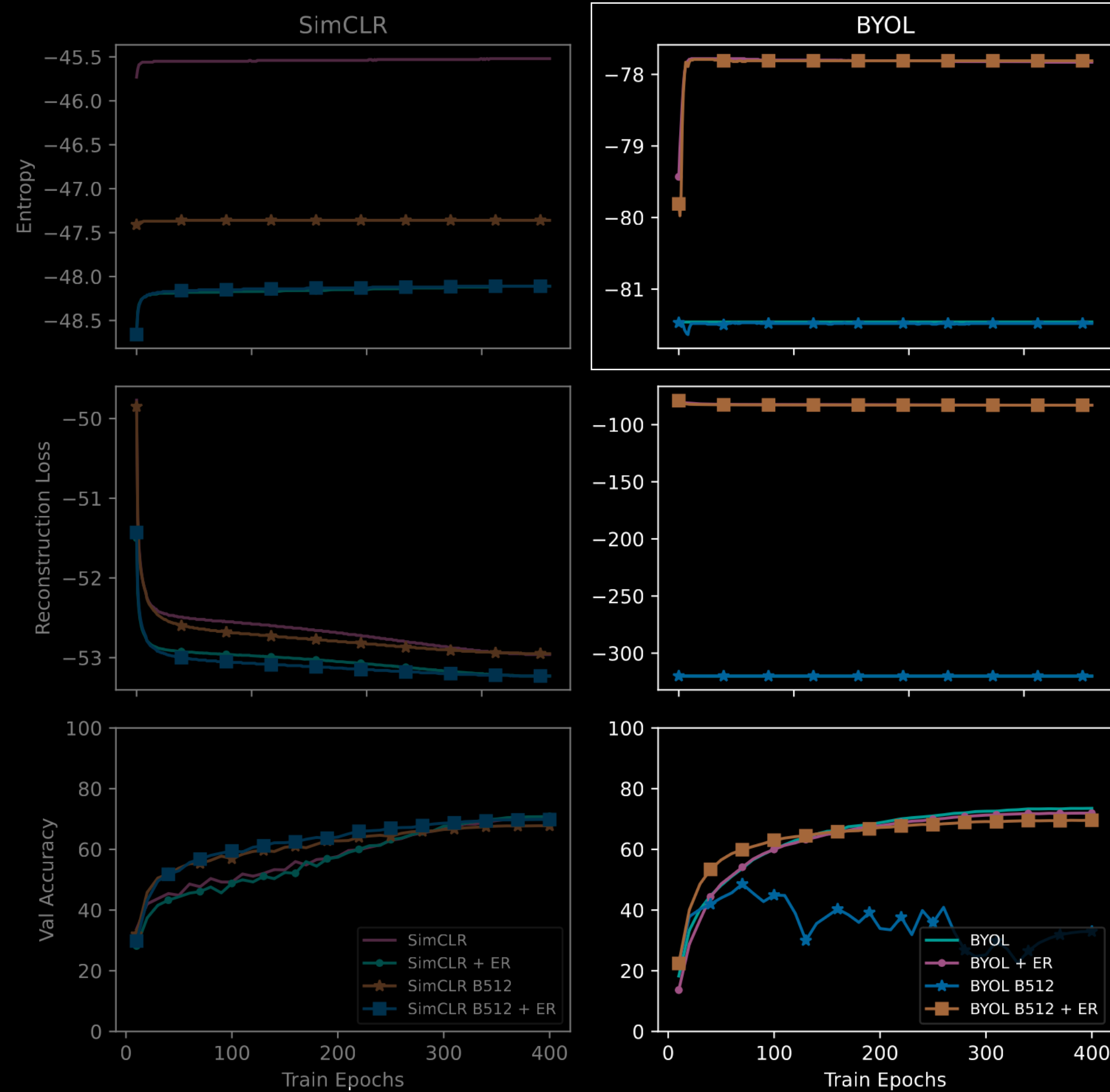
Empirical results: We can add entropy optimization explicitly



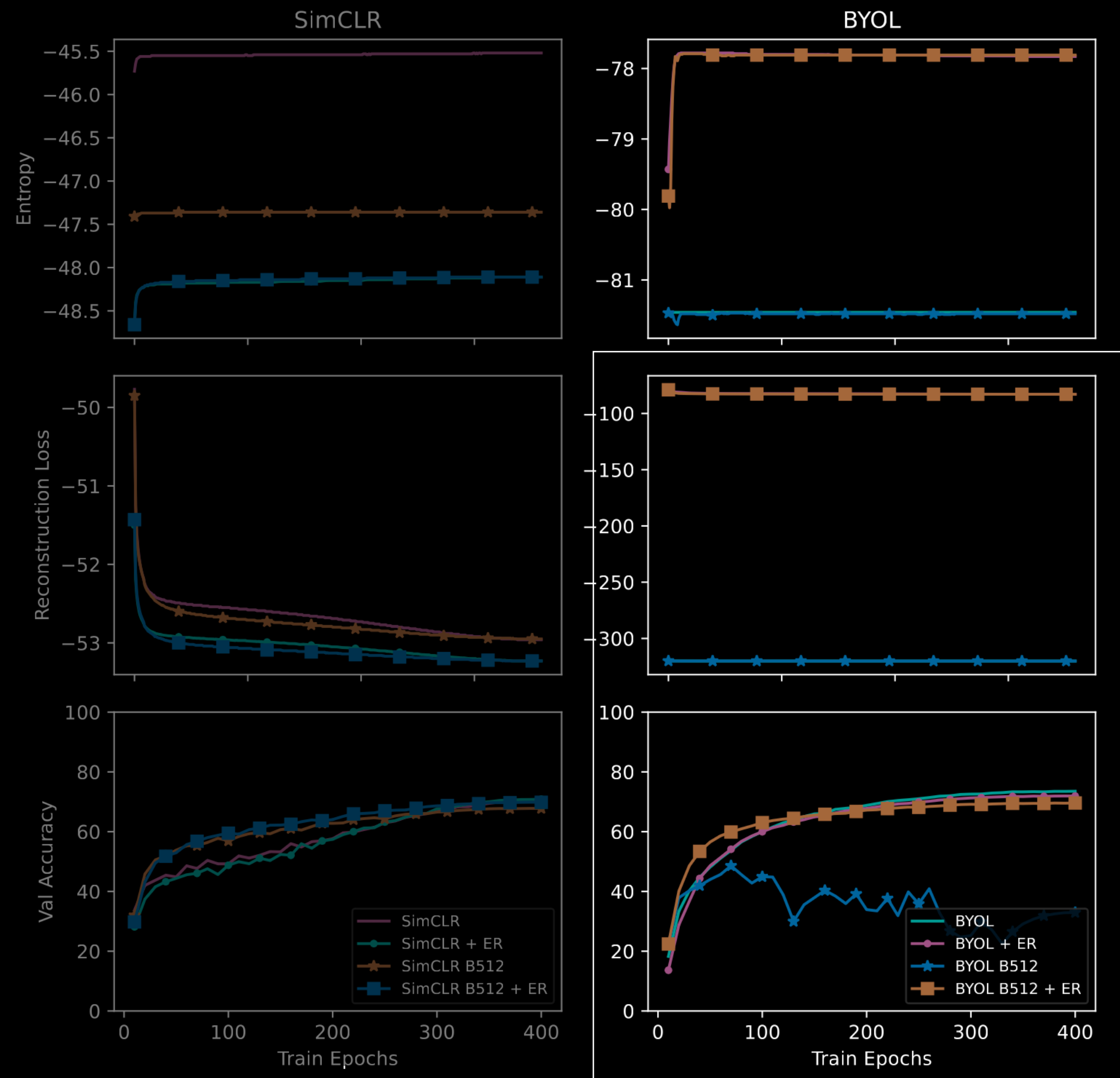
Empirical results: We can add entropy optimization explicitly



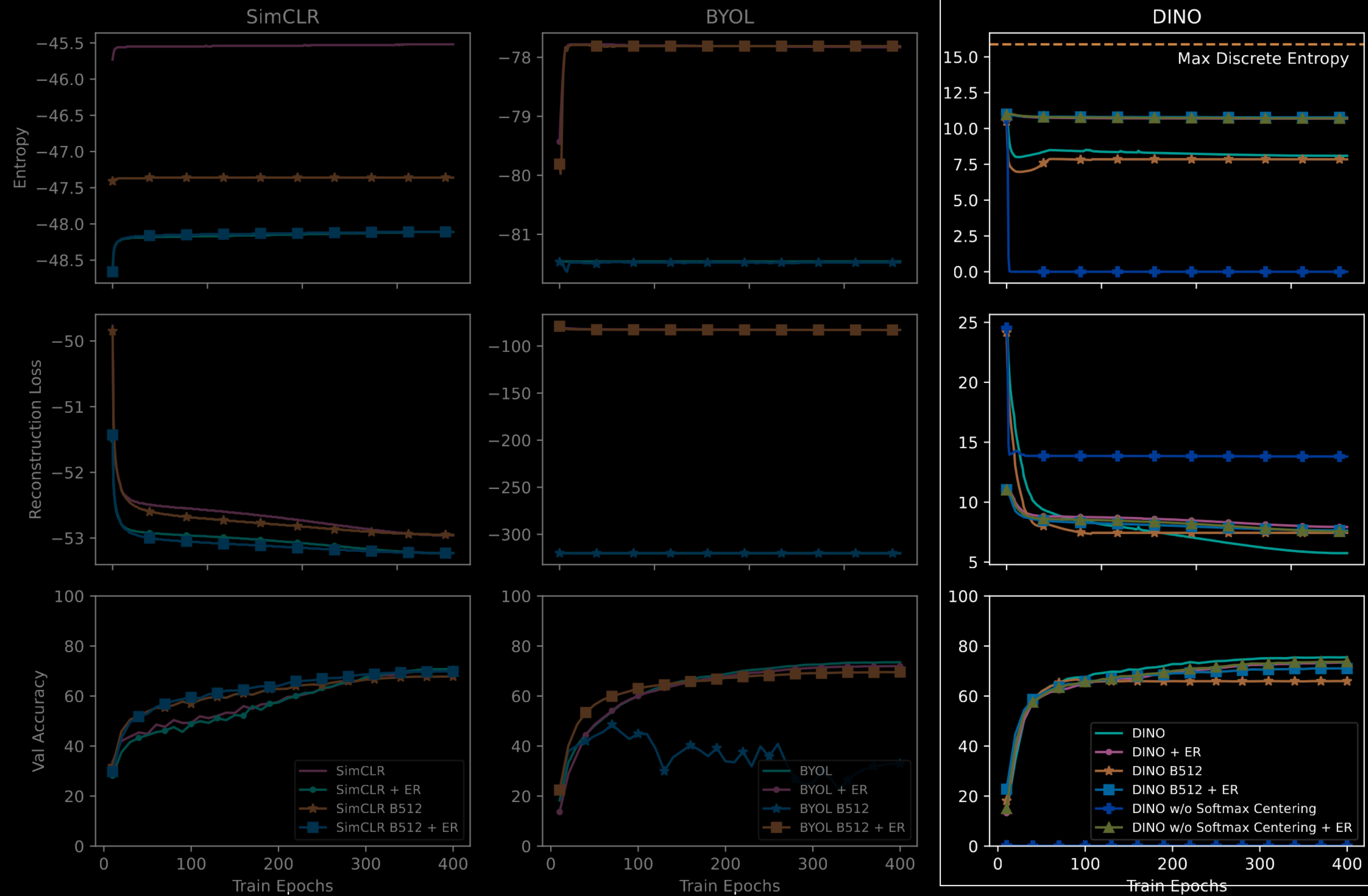
Empirical results: We can add entropy optimization explicitly



Empirical results: We can add entropy optimization explicitly



Empirical results: We can add entropy optimization explicitly



Empirical results: We can add entropy optimization explicitly

