UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Jesimon Barreto Santos

VESSA: Video-based Efficient Self-Supervised Adaptation for visual foundation models

Belo Horizonte
2025

Jesimon Barreto Santos

VESSA: Video-based Efficient Self-Supervised Adaptation for visual foundation models

Final Version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: William Robson Schwartz
Co-Advisor: André Filgueiras de Araujo

Belo Horizonte
2025

Santos, Jesimon Barreto.

S237v VESSA: [recurso eletrônico] Video-based Efficient Self-Supervised Adaptation for visual foundation models / Jesimon Barreto Santos – 2025.
1 recurso online (71 f. il., color.) : pdf.

Orientador: William Robson Schwartz.
Coorientador: André de Filgueiras Araújo.

Dissertação (Mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Referências: f. 60-65.

1. Computação – Teses. 2. Visão por computador – Teses. 3. Processamento de imagens – Teses. I. Schwartz, William Robson. II. Araújo, André de Filgueiras. III. Universidade Federal de Minas Gerais Instituto de Ciências Exatas, Departamento de Ciência da Computação. IV. Título.

CDU 519.6*82.10(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais – ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

VESSA: VIDEO-BASED EFFICIENT SELF-SUPERVISED ADAPTATION FOR VISUAL FOUNDATION MODELS

JESIMON BARRETO SANTOS

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores(a):

Prof. William Robson Schwartz - Orientador
Departamento de Ciência da Computação - UFMG

Doutor André Araújo - Coorientador
DeepMind - Google

Prof. Pedro Olmo Stancioli Vaz de Melo
Departamento de Ciência da Computação - UFMG

Prof. David Menotti Gomes
Departamento de Informática - UFPR

Belo Horizonte, 22 de julho de 2025.

Documento assinado eletronicamente por **William Robson Schwartz**, **Professor do**

Magistério Superior, em 28/07/2025, às 14:46, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **David Menotti Gomes**, **Usuário Externo**, em 28/07/2025, às 16:41, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Pedro Olmo Stancioli Vaz de Melo**, **Professor do Magistério Superior**, em 10/09/2025, às 17:23, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4393083** e o código CRC **9BD059EE**.

Referência: Processo nº 23072.244200/2025-91 SEI nº 4393083

Documento assinado digitalmente
ANDRE FILGUEIRAS DE ARAUJO
Data: 24/10/2025 15:25:30-0300
Verifique em https://validar.iti.gov.br

Acknowledgments

I would like to thank the *National Council for Scientific and Technological Development* – CNPq (Grant 312565/2023-2) and the *Coordination for the Improvement of Higher Education Personnel* (CAPES) for their invaluable support throughout this research. I am also deeply thankful to *Google*, which provided essential financial support and access to TPU resources, enabling the execution of large-scale experiments. Additionally, I acknowledge the *Google Academic Research Grants* program for providing generous credits to access TPUs and various tools on Google Cloud, which were instrumental in carrying out the experiments and analyses presented in this dissertation. Finally, I extend my gratitude to all those who, directly or indirectly, contributed to the development of this work through their guidance, collaboration, and encouragement.

"*To understand images, we must first learn to represent them.*"
(Alexei A. Efros)

Resumo

Modelos fundacionais têm impulsionado avanços em visão computacional, alcançando alto desempenho em diversas tarefas por meio de pré-treinamento em larga escala e ajuste supervisionado. No entanto, esses modelos podem apresentar desempenho insatisfatório em domínios com mudanças de distribuição e escassez de rótulos, onde o ajuste supervisionado não é viável. Embora a continuação do aprendizado auto-supervisionado seja comum em modelos de linguagem generativos, essa abordagem ainda não mostrou eficácia em modelos de codificação centrados em visão. Para enfrentar esse desafio, propomos uma nova formulação de ajuste fino auto-supervisionado para modelos fundacionais visuais, na qual o modelo é adaptado a um novo domínio sem necessidade de anotações, utilizando apenas vídeos curtos centrados em objetos.

Neste trabalho, é proposta a VESSA: **V**ideo-based **E**fficient **S**elf-**S**upervised **A**daptation for visual foundation models. A técnica de treinamento VESSA baseia-se em um paradigma de auto-destilação, no qual é essencial ajustar cuidadosamente as cabeças de predição e utilizar técnicas de adaptação eficientes em parâmetros — caso contrário, o modelo pode esquecer rapidamente o conhecimento prévio. VESSA se beneficia significativamente de observações de objetos em diferentes quadros de vídeo, aprendendo de forma eficiente a robustez frente a variações nas condições de captura, sem necessidade de rótulos.

Por meio de experimentos abrangentes com três modelos fundacionais de visão em dois conjuntos de dados, VESSA demonstra melhorias consistentes em tarefas de classificação, superando os modelos base e métodos anteriores de adaptação. Os conjuntos de dados utilizados nos experimentos foram CO3D e MVImageNet, e os modelos fundationais visuais avaliados incluem DINO, DINOv2 e TIPS.

Palavras-chave: modelos fundacionais visuais; ajuste fino auto-supervisionado; adaptação baseada em vídeo; ajuste eficiente em parâmetros.

Abstract

Foundation models have advanced computer vision by enabling strong performance across diverse tasks through large-scale pretraining and supervised fine-tuning. However, they may underperform in domains with distribution shifts and scarce labels, where supervised fine-tuning may be infeasible. While continued self-supervised learning for model adaptation is common for generative language models, this strategy has not proven effective for vision-centric encoder models. To address this challenge, we introduce a novel formulation of self-supervised fine-tuning for vision foundation models, where the model is adapted to a new domain without requiring annotations, leveraging only short object-centric videos.

In this work, we propose VESSA:**V**ideo-based **E**fficient **S**elf-**S**upervised **A**daptation for visual foundation models. VESSA's training technique is based on a self-distillation paradigm, where it is critical to carefully tune prediction heads and deploy parameter-efficient adaptation techniques – otherwise, the model may quickly forget its pretrained knowledge and reach a degraded state. VESSA benefits significantly from object observations sourced from different frames in a video, efficiently learning robustness to varied capture conditions, without the need of annotations.

Through comprehensive experiments with 3 vision foundation models on 2 datasets, VESSA demonstrates consistent improvements in downstream classification tasks, compared to the base models and previous adaptation methods. The datasets used in our experiments are CO3D and MVImageNet, and the visual foundation models evaluated include DINO, DINOv2, and TIPS.

Keywords: vision foundation models; self-supervised fine-tuning; video-based adaptation; parameter-efficient tuning.

List of Figures

List of Tables

List of Abbreviations and Acronyms

CL	*Contrastive Learning*
CNN	*Convolutional Neural Network*
CV	*Computer Vision*
DINO	*Self-Distillation with No Labels*
EMA	*Exponential Moving Average*
FM	*Foundation Model*
MAE	*Masked Autoencoder*
NLP	*Natural Language Processing*
SSL	*Self-Supervised Learning*
ViT	*Vision Transformer*
VESSA	*Video-based Efficient Self-Supervised Adaptation*
VFM	*Visual Foundation Model*
LoRA	*Low-Rank Adaptation*
KNN	*k-Nearest Neighbors*

Contents

Chapter 1

Introduction

Computer vision (CV) plays a pivotal role in enabling machines to perceive, interpret, and interact with the visual world. From everyday applications such as facial recognition on smartphones, autonomous driving, and medical imaging, to industrial automation and surveillance, the ability to extract meaningful information from visual data has become increasingly indispensable. Illustrative examples of these applications are presented in Figure 1.1. Despite this progress, visual understanding remains a fundamentally complex task due to the high dimensionality of images, variability in object appearance, occlusions, dynamic environmental conditions, and the diversity of application domains. Traditionally, these challenges were addressed in a task-specific manner, with models being designed and optimized for narrowly defined problems. This fragmented approach resulted in highly specialized algorithms or models that often lacked generalization capabilities across different visual tasks and domains.

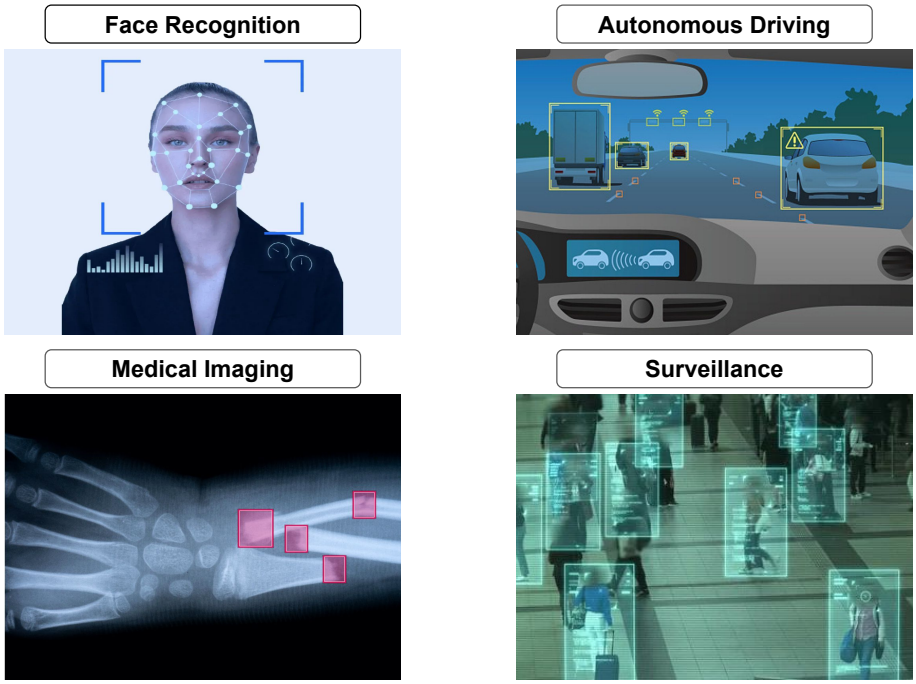

Figure 1.1: Examples of computer vision applications in daily life: facial recognition, autonomous driving, medical imaging, and surveillance.

In recent years, the concept of *foundation models* has emerged as a central paradigm in machine learning, particularly in natural language processing (NLP) and computer vision. These models are characterized by their large scale, general-purpose design, and ability to be adapted across a wide range of downstream tasks [Awais et al., 2025]. A foundation model is typically trained using vast amounts of data—often sourced from diverse and heterogeneous distributions—combined with self-supervised or weakly supervised objectives. The aim of this process is not to optimize performance for any specific task, but to endow the model with a general understanding of the domain, which can then be fine-tuned or specialized later. In the context of computer vision, this typically means training a model to extract rich, reusable visual representations from large collections of images. As a result, once pre-trained, the foundation model can be adapted to a variety of applications such as classification, detection, segmentation, or retrieval. The success of these models lies in their capacity to generalize from pre-training data to unseen downstream tasks with relatively little supervision [Bommasani et al., 2021, Han et al., 2022].

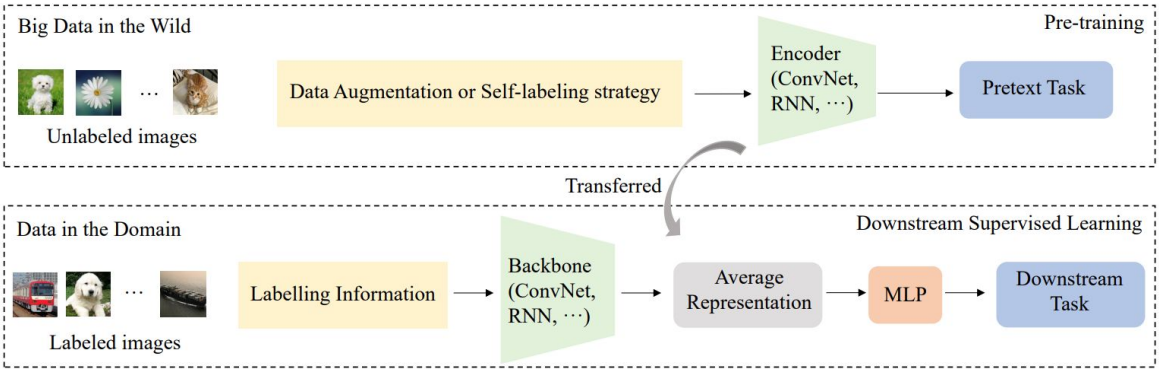

Figure 1.2: Overview of the vision foundation model pipeline: pre-training using large-scale self-supervised learning, followed by domain-specific fine-tuning for downstream tasks.

Source: Zhou et al. [2024]

The construction of a vision foundation model involves training a deep neural network—often with hundreds of millions or even billions of parameters—on a large image dataset. Instead of relying on human-annotated labels, the model is typically optimized using self-supervised learning techniques. These approaches exploit the inherent structure

in the data to create proxy tasks that guide representation learning. For example, contrastive learning methods [Jaiswal et al., 2020] train the model to bring representations of different augmentations of the same image closer together while pushing apart representations from different images. Other strategies, such as masked image modeling [Li et al., 2023] and self-distillation [Zhang et al., 2021], encourage the model to reconstruct missing parts of the input or align representations across network layers. The result of this training process is a model that encodes general visual features, which can then be reused in new domains or tasks with minimal supervision. Figure 1.2 provides a high-level schematic of this pre-training and downstream pipeline.

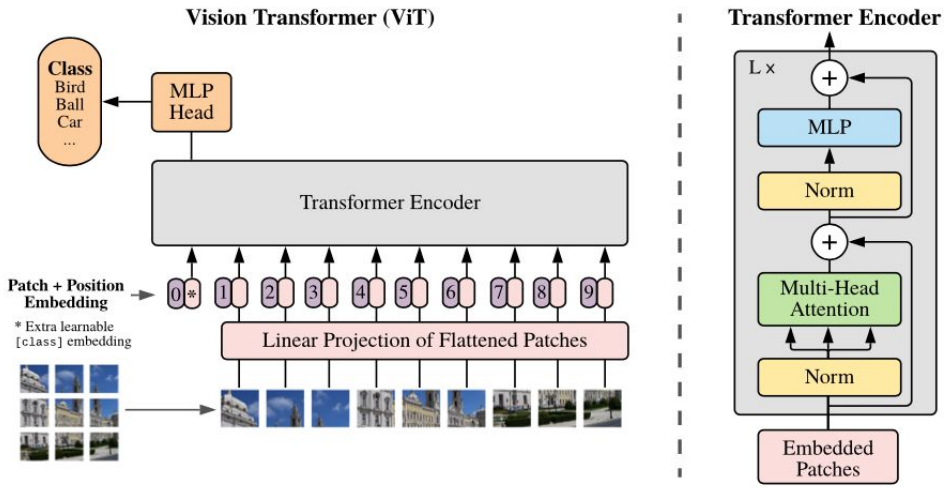

Figure 1.3: Illustration of the Vision Transformer (ViT) architecture. The image is divided into patches, which are embedded and processed through multiple layers of self-attention and feedforward networks.

Source: Dosovitskiy et al. [2021]

A critical technological advancement that has enabled the success of modern foundation models is the *transformer* architecture [Zhang et al., 2023]. Originally proposed for natural language processing, transformers are based on the self-attention mechanism, which allows the model to dynamically weigh relationships between all elements in the input. Unlike convolutional neural networks (CNNs), which rely on spatial locality and fixed receptive fields, transformers can model long-range dependencies and capture complex global patterns [Lu et al., 2021]. This makes them particularly well-suited for tasks where understanding the global context of the input is important. The flexibility and scalability of the transformer architecture have led to its widespread adoption in vision tasks through adaptations such as the Vision Transformer (ViT) [Dosovitskiy et al., 2021]. These models treat an image as a sequence of non-overlapping patches, embedding them into tokens that can be processed similarly to words in a sentence.

Transformer-based vision models have demonstrated competitive or superior performance compared to traditional CNNs across a wide variety of benchmarks [Lu et al., 2021]. Their modular design also facilitates the incorporation of advanced techniques such as hierarchical processing, cross-modal learning, and efficient fine-tuning [Khan et al., 2022]. Furthermore, self-supervised pre-training has been especially effective when paired with transformers, enabling models like DINO [Caron et al., 2021], MAE [He et al., 2022] and DINOv2 [Oquab et al., 2023] to learn high-quality representations without requiring any human annotations. Figure 1.3 illustrates the general architecture of a Vision Transformer, highlighting the patch embedding, positional encoding, multi-head self-attention blocks, and classification head.

Despite the remarkable success of transformer-based foundation models across a broad spectrum of visual tasks, their general-purpose nature often falls short in specialized domains that exhibit significant distributional shifts from the data used during pretraining. In such cases, performance can degrade substantially, motivating the need for adaptation strategies tailored to the target domain. The predominant solution in the literature has been to apply supervised fine-tuning, which leverages labeled examples to align the model with the new domain-specific distribution. While this approach has proven effective in many contexts, it implicitly assumes the availability of annotated data—a condition that is often unrealistic in real-world scenarios due to cost, scale, or domain complexity. This work addresses the methodological gap that arises in such settings by exploring alternative strategies for adapting vision foundation models in the absence of labels.

1.1 Motivation

Visual foundation models trained with self-supervised learning on large image datasets have become a powerful tool for a wide range of computer vision tasks [Gui et al., 2024, Awais et al., 2025]. Techniques such as contrastive learning and self-distillation allow these models to learn high-quality visual representations without manual labels [Caron et al., 2021, Oquab et al., 2023]. Despite their generality, performance can suffer when applied to specialized domains with different characteristics from the pre-training data. For this reason, after the VFM is pre-trained, fine-tuning is commonly employed before applying it to downstream tasks. Supervised fine-tuning, in particular, has been the dominant approach, with impressive results across a variety of datasets and applications [Awais et al., 2025, Han et al., 2022] such as remote sensing [He et al., 2023, Zou et al., 2024], medical imaging [Cui et al., 2024, Baharoon et al., 2023] and place recognition

[Izquierdo and Civera, 2024, Lu et al., 2024]. These successes demonstrate the adaptability of pre-trained models, but also highlight their reliance on labeled data, which can be expensive or impractical to obtain in many real-world scenarios.

Significant challenges may arise in scenarios where labeled data are unavailable for supervised fine-tuning. Despite its potential in cases where collecting annotations is costly, time-consuming, or even infeasible, unsupervised fine-tuning for foundation models remains largely underexplored in vision [Dong et al., 2025, Chen et al., 2023]. By contrast, the NLP community has long adopted unsupervised fine-tuning as a standard method to specialize large language models to new data distributions, typically via continued pretraining on unlabeled in-domain text [Gururangan et al., 2020, Han et al., 2020, Liu et al., 2021]. While this strategy has proven successful for generative language models, its adaptation to visual data remains an open and challenging problem. For this reason, a few natural questions arise: how can we adapt a vision pre-trained model to a specific context without supervision? What forms of unlabeled visual data are best suited for adapting vision foundation models to new data distributions? What type of learning technique can effectively adapt pre-trained visual representations under the constraints in this scenario?

1.2 Objectives

The primary objective of this work is to propose a novel technique for adapting visual foundation models (VFMs) without requiring labeled samples from the target domain, aiming to capture better discriminative representations in specialized settings where pre-trained models often fail to deliver satisfactory results. While VFMs demonstrate impressive performance across a wide range of generic computer vision tasks, their general-purpose training leads to limitations when applied to datasets that significantly diverge from the large-scale, natural image distributions used during pre-training. To address this issue, this work investigates and designs a self-supervised learning (SSL) adaptation strategy capable of enhancing the feature representations of VFMs in such domains, improving performance without relying on any form of manual annotation. In this work, we focus on short videos with the distinctive characteristic of being object-centric, which allows the model to leverage consistent visual context while learning from variations in object appearance and viewpoint.

The primary objective of this research can be divided into three specific goals. First, to understand the behavior and performance of visual foundation models, including an empirical analysis of common adaptation strategies—such as continued self-supervised learning with domain-specific data—and to demonstrate that these models often underper-

form in such scenarios or that naive continued training can be ineffective or even detrimental. Second, to explore the use of domain-specific visual patterns—such as object-centric motion in videos—as a source of supervisory signal capable of enhancing representation quality, considering that videos inherently provide richer and more diverse information about objects, including appearance variation across viewpoints and temporal coherence, which can significantly benefit the learning and adaptation of models to new domains. Third, to develop and fine-tune the technical components of the adaptation pipeline, including data preprocessing, data transformations, loss functions, and hyperparameter configurations, in order to achieve measurable improvements in downstream task performance; emphasizing that adaptation plays a crucial role in this context, as the learning process involves transferring knowledge from video data to image-based tasks, with the primary objective being domain adaptation rather than training a model from scratch. Together, these steps aim to validate the effectiveness of the proposed adaptation technique.

1.3 Contributions

The contribution proposed in this work is formulated as a direct response to the research questions presented at the end of Section 1.1. To answer the research questions, we propose VESSA (**V**ideo-based **E**fficient **S**elf-**S**upervised **A**daptation), a self-supervised fine-tuning method for VFMs that is both simple and effective, leveraging only short object-centric videos. We evaluate our approach using short videos with a clear focus on the central object, captured against standardized backgrounds; examples of characteristic video frames can be seen in Fig. 5.2. A conceptual overview of the application of our model is presented in Fig. 1.4. VESSA employs a self-distillation training algorithm with critical adaptations to make it work in a fine-tuning setup. We show that a naive application of self-distillation to the fine-tuning stage may lead to a degraded model state, but this can be avoided with the careful adjustments proposed in this work.

In particular, we introduce a training schedule which adjusts the self-distillation prediction head before unfreezing the rest of the model. Then, an efficient method is used to gently tune the backbone parameters towards the new domain without disrupting the encoded pre-trained knowledge, also leveraging uncertainty weighting to prioritize harder training examples. Finally, we propose to source observations of target objects in the new domain from short videos, which are easy to capture and require no labeling, but enhance the model performance significantly.

Experimentally, we leverage three existing foundation models and two downstream

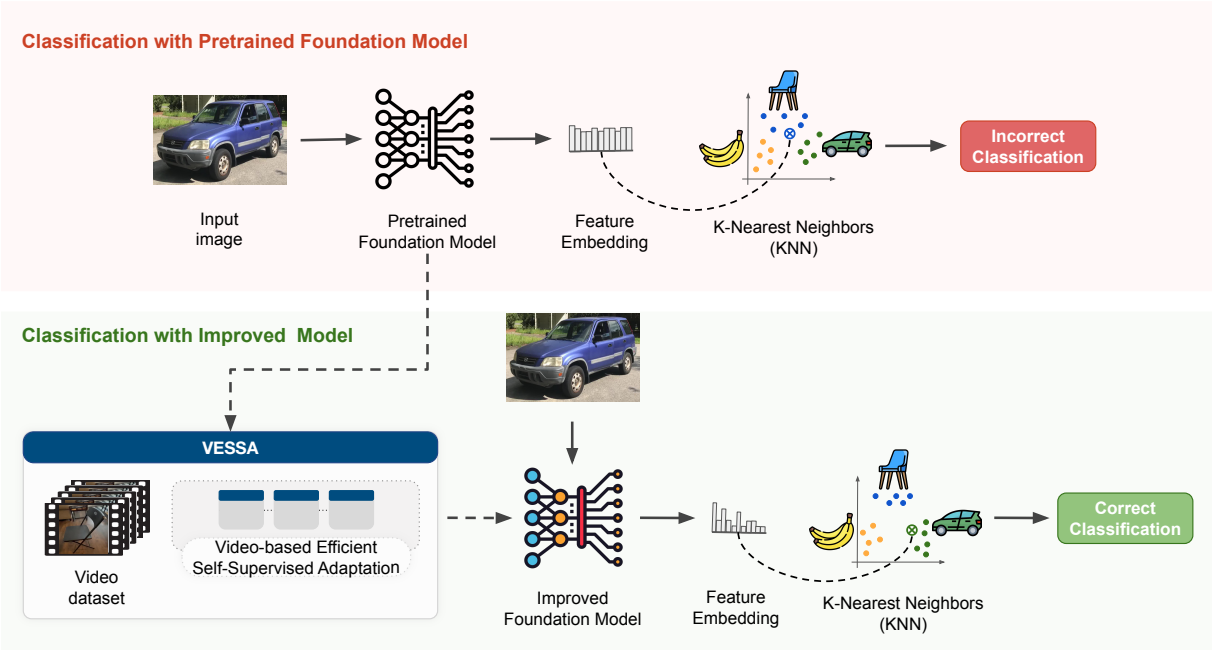

Figure 1.4: The **VESSA**, a novel and efficient method for adapting vision foundation models using self-supervised fine-tuning with videos. Starting from a pretrained foundation model applied to a classification problem in a target domain, VESSA adapts the model without using labels by leveraging simple, object-centric videos. The resulting model learns improved representations that better structure the feature space in the target domain, boosting downstream classification accuracy.

classification applications to comprehensively assess the proposed VESSA technique. Our results demonstrate that the proposed video-based self-supervised fine-tuning significantly outperforms base foundation models or other fine-tuning strategies.

The main contribution of this dissertation resulted in the publication of a paper at *NeurIPS 2025 — The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, one of the most prestigious venues in the field. The paper, entitled *"VESSA: Video-based objEct-centric Self-Supervised Adaptation for Visual Foundation Models"*[1], constitutes the core scientific contribution of this thesis and is hereafter referred to as Barreto et al. [2025]. This work introduces an efficient and domain-adaptive training framework for visual foundation models, leveraging object-centric multi-view videos for improved generalization across domains.

[1]Barreto, J., Caetano, C., Araujo, André, Schwartz, W. R. (2025). *VESSA: Video-based objEct-centric Self-Supervised Adaptation for Visual Foundation Models.* In *Proceedings of NeurIPS 2025 — The Thirty-Ninth Annual Conference on Neural Information Processing Systems.*

1.4 Work Organization

The remainder of this thesis is structured as follows. Chapter 2 presents the related work and is organized into three main topics: self-supervised visual foundation models, task-adaptive fine-tuning, and video-to-image knowledge transfer. Chapter 3 introduces the theoretical background that supports this research, focusing on two key components: DINO [Caron et al., 2021], a self-distillation framework for vision transformers, and LoRA [Hu et al., 2021], a parameter-efficient fine-tuning technique. Chapter 4 describes the proposed approach, titled *VESSA* (Video-based Efficient Self-Supervised Adaptation), and is divided into two sections: the main adaptation strategy and a discussion of critical optimization considerations necessary for effective model adaptation. Chapter 5 reports the experimental study and is structured into the experimental setup and the analysis of results. Finally, Chapter 6 presents the main conclusions of this work and is divided into two sections discussing the limitations of the proposed approach and directions for future research.

Chapter 2

Related Work

Recent advances in visual foundation models have reshaped the landscape of computer vision by enabling scalable, general-purpose representations trained on massive datasets. To tailor these representations to specific downstream tasks, task-adaptive fine-tuning strategies have emerged as a solution for many applications, aiming to bridge the gap between foundation model generality and task-specific performance. Complementary to this, approaches in video-to-image knowledge transfer explore how temporal and multimodal supervision in video models can be distilled into stronger static image representations. In this chapter, we provide a structural overview of these areas, highlighting their connections to our proposed formulation and identifying key gaps our method addresses.

2.1 Self-supervised Visual Foundation Models

Self-supervised learning (SSL) is a learning paradigm in which models are trained using automatically generated supervision signals derived from the data itself, without relying on manual annotations. The core idea of SSL is to design pretext tasks—auxiliary learning objectives that require the model to predict certain aspects of the input data. By solving these tasks, the model learns to extract semantically meaningful and generalizable representations, which can be transferred to downstream tasks such as classification, detection, or segmentation with minimal or no additional supervision.

SSL has led to the development of a wide range of methods that have been essential for training Vision Foundation Models (VFMs) without relying on human supervision. These methods are commonly categorized according to their pretext tasks, which define the form of supervision used during pretraining. The principal categories include: classical pretext tasks, contrastive learning, masked image modeling, clustering-based learning, and self-distillation. Each category introduces distinct inductive biases and involves specific trade-offs in terms of scalability, representation quality, and effectiveness in domain-specific contexts. Illustrative examples of several pretext tasks are shown in Figure 2.1.

While these tasks are not necessarily the most representative or categorized, the figure provides a visual overview of many common approaches. Each task is designed according to the visual problem it aims to address during pretraining. In this context, the DNN represents a generic deep neural network that is trained to solve the corresponding pretext task and to learn transferable visual representations.

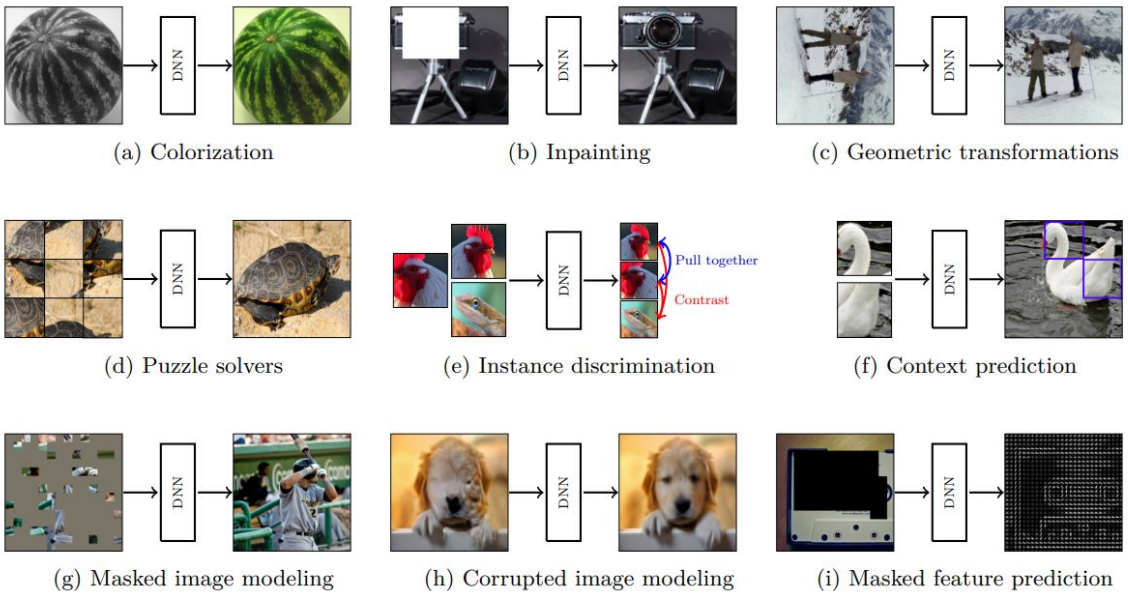

Figure 2.1: Illustrations of various image-based pretext tasks used in self-supervised learning, highlighting the underlying training objectives each method is designed to solve.

Source: Ozbulak et al. [2023]

Classical pretext tasks laid the foundation for modern SSL by defining simple objectives that encourage visual understanding without annotated data. These include tasks such as rotation prediction [Gidaris et al., 2018], jigsaw puzzle solving [Noroozi and Favaro, 2016], and image colorization [Zhang et al., 2016]. While their downstream performance was limited compared to modern approaches, they demonstrated the feasibility of unsupervised representation learning and informed key aspects of later SSL frameworks, such as the role of structured transformations and reconstruction objectives. Many of these ideas would later re-emerge in masked modeling and hybrid architectures.

Contrastive learning methods advanced SSL by introducing objectives based on instance discrimination—pulling together representations of different augmented views of the same image while pushing apart views from different images. SimCLR [Chen et al., 2020] and MoCo [He et al., 2020] are prominent examples that leverage large batch sizes or memory banks to provide sufficient negative samples. These methods demonstrated impressive performance but exhibit sensitivity to batch composition and augmentation diversity. Later methods like BYOL [Grill et al., 2020] and SimSiam [Chen

and He, 2021] removed the need for explicit negatives using asymmetric architectures and stop-gradient techniques to prevent collapse, though often requiring careful tuning of optimization dynamics and hyperparameters.

Masked image modeling (MIM) presents an alternative paradigm where models are trained to reconstruct occluded parts of the input image. MAE [He et al., 2022] introduced a lightweight decoder and aggressive masking to enable scalable training on high-resolution images. BEiT [Bao et al., 2022], by contrast, adopts a discrete token prediction objective, leveraging a pretrained tokenizer to transform image patches into token indices. MIM methods are well-suited to learning fine-grained local representations and have become strong initializations for vision transformers. However, they often require large-scale training data and may underperform in scenarios requiring global semantic consistency unless complemented with auxiliary tasks or large models.

Clustering-based methods such as DeepCluster [Caron et al., 2018], SeLa [Asano et al., 2020], and SwAV [Caron et al., 2020] use pseudo-labels derived from unsupervised clustering to train the model. These methods promote consistency in cluster assignments across augmentations and encourage semantic organization in the embedding space. SwAV, in particular, improves scalability by introducing online clustering and multi-crop augmentations. Despite their merits, clustering-based methods often require periodic offline computations or sophisticated optimization procedures, which can complicate their deployment in online or data-limited settings.

Among all paradigms, self-distillation without labels has emerged as particularly effective and scalable. DINO [Caron et al., 2021] introduced a teacher-student architecture where the teacher, updated via an exponential moving average, provides soft targets for the student model. This approach eliminates the need for negative samples and fosters semantic representations that emerge naturally from enforcing consistency between multi-view outputs. iBOT [Zhou et al., 2022] expands this paradigm by incorporating masked token prediction within the distillation framework, blurring the line between MIM and distillation. DINOv2 [Oquab et al., 2023] significantly improves upon DINO by refining data curation, training pipelines, and architectural scaling, resulting in highly robust foundation models with state-of-the-art performance in zero-shot, linear probing, and dense prediction tasks. Distillation methods offer a compelling trade-off between architectural simplicity, training stability, and generalization capacity, making them particularly attractive for adapting models in settings with limited or no labeled data.

This work builds upon the self-distillation family, particularly the DINO paradigm, and adapts it to exploit temporal coherence in videos. While most visual foundation models are pretrained on still images, we propose leveraging videos to provide temporally diverse yet semantically consistent views of the same object, offering richer learning signals without supervision. This approach addresses a key limitation of existing methods that do not fully utilize temporal redundancy in sequential data. Although our method

is grounded in the self-distillation framework, it can also serve as a post-hoc adaptation strategy for foundation models initially pretrained using other SSL paradigms, such as contrastive learning or masked image modeling. By integrating video-based multi-view learning into the distillation framework, we aim to enhance the adaptability of foundation models to specialized domains where annotations are scarce but temporal data is abundant.

2.2 Task-Adaptive Fine-Tuning

Task-adaptive pretraining, also referred to as continual pretraining, consists of methods that extend self-supervised learning beyond the initial pretraining stage by using different data distributions. The primary goal is to adapt foundation models to new domains without requiring manual annotations. This approach has been instrumental in enabling large foundation models to specialize for specific downstream domains, particularly in NLP [Gururangan et al., 2020, Han et al., 2020, Liu et al., 2021]. In this paradigm, a pretrained model is further refined using unlabeled data from a target domain prior to task-specific fine-tuning. Although originally developed in the context of NLP, the core ideas have recently been applied to computer vision, where distribution shifts and the scarcity of labeled data continue to pose significant challenges.

In the vision domain, most task-adaptive approaches have focused on adapting vision foundation models (VFMs) via supervised pipelines. Techniques such as AdaptFormer [Chen et al., 2022] and Visual Prompt Tuning [Jia et al., 2022] allow parameter-efficient transfer to new tasks, but still rely on labeled data for fine-tuning. *AdaptFormer*, for instance, introduces a lightweight module composed of two fully connected layers, a non-linear activation function, and a scaling factor, which is added in parallel to the feed-forward network of the Vision Transformer. This design enables efficient adaptation using a very small number of trainable parameters. However, such methods are task-specific and require careful engineering of adaptation layers or prompts, which can become complex and brittle when scaling across diverse domains.

To overcome the reliance on supervision, recent works have explored task-adaptive pretraining through more self-supervised and domain-specific strategies. For example, Scheibenreif et al. [2024] propose a parameter-efficient framework for adapting remote sensing foundation models to new data modalities using *Scaled Low-Rank (SLR) adapters*. These lightweight modules are optimized through self-supervised learning on unlabeled data from the target domain, while keeping the backbone frozen. This allows for efficient adaptation without discarding prior knowledge, and proves particularly effective in

low-resource and few-shot settings. In a complementary direction, Mendieta et al. [2023] introduce the *Geospatial Foundation Model (GFM)*, which builds on general-purpose ImageNet models and extends them via continual pretraining on geospatial imagery. Their method uses a multi-objective optimization strategy combining masked image modeling and teacher-student self-distillation, enabling the model to learn domain-relevant features without explicit supervision. While both approaches advance the field by leveraging self-supervised adaptation to specialize vision models for the geospatial domain, they still rely on supervised fine-tuning for downstream tasks and often require assembling new foundation models tailored to the target domain.

ExPLoRA [Khanna et al., 2024] takes this one step further by incorporating parameter-efficient techniques such as LoRA [Hu et al., 2021] into the continual pretraining of VFMs like DINOv2 [Oquab et al., 2023] and MAE [He et al., 2022], specifically for satellite imagery. The method avoids costly full-domain pretraining by initializing a vision transformer (ViT) with weights from a generalist model and selectively unfreezing one or two transformer blocks. It then applies LoRA to the query and value projections of attention layers in the remaining frozen blocks, while also unfreezing normalization layers to enhance adaptability. Using the same self-supervised objective as the source model (e.g., DINO or MAE), ExPLoRA continues training on unlabeled data from the target domain to learn a structured, low-rank update Δ_T, resulting in a new adapted model $W_T^* = W_S + \Delta_T$.

This approach enables efficient domain adaptation using less than 10% of the trainable parameters and demonstrates strong performance across various satellite benchmarks such as fMoW, including RGB, temporal, and multi-spectral imagery. Additionally, ExPLoRA generalizes well to non-satellite datasets, outperforming full fine-tuning approaches on domains like wildlife, agriculture, and medical imaging in the WILDS benchmark [Koh et al., 2021]. While ExPLoRA reuses the architecture and training heads of base models, it still depends on task-specific fine-tuning for final performance. In contrast, our method departs from this dependency by adapting the pretrained DINO architecture [Caron et al., 2021] directly to a new domain without constructing a new foundation model or relying on any labeled data.

Our approach introduces a novel form of task-adaptive fine-tuning, in which the adaptation is performed entirely through self-supervised learning. The core distinction of our method lies in a new loss function that modifies the learning dynamics by encouraging the extraction of fine-grained details specific to the target domain. In addition, we propose a carefully designed parameter configuration that applies LoRA in a domain-adaptive yet efficient manner, enabling targeted updates while preserving the compactness of the model. Importantly, our framework establishes a consistent adaptation strategy that can be applied independently of the original pretraining method. A key differentiator of our approach is the use of knowledge derived from videos to guide the adaptation of still

images, as detailed in the next section.

2.3 Video to Image Knowledge Transfer

Videos provide a natural source of supervision for representation learning due to their rich spatio-temporal structure. Temporal continuity, object persistence, and viewpoint variation across frames offer implicit cues that can be exploited for learning semantics without labels [Agrawal et al., 2015, Wang and Gupta, 2015, Pathak et al., 2017, Goroshin et al., 2015, Misra et al., 2016, Kulkarni et al., 2019]. Unlike static images, videos capture natural object deformations, scale changes, occlusions, and transformations in context, making them an attractive modality for self-supervised learning.

Recent research has increasingly emphasized the potential of leveraging video-based supervision to enhance image-level representations. Videos naturally encode rich supervisory signals such as object permanence, motion cues, and temporal consistency, which are often absent in still image datasets. Consequently, transferring knowledge from videos to images has emerged as a promising direction for improving the generalization and robustness of visual models, especially in self-supervised scenarios [Aubret et al., 2023].

Despite this potential, many existing approaches suffer from practical limitations due to their reliance on intricate multi-frame architectures and hybrid training objectives. For instance, ViC-MAE [Hernandez et al., 2024] combines masked image modeling with contrastive learning by treating short video clips as temporally coherent augmentations. While this strategy enables ViC-MAE to outperform previous video-to-image transfer models such as OmniMAE, it requires a carefully balanced design that samples frames at large temporal gaps (approximately 1.06 seconds) and applies strong augmentations to simulate diverse views. Although more efficient than frame-dense models like ST-MAE, ViC-MAE still incurs higher computational overhead than traditional MAE approaches due to its dual objectives and increased token handling. Nevertheless, it demonstrates that integrating temporal diversity into self-supervised learning pipelines significantly improves downstream performance on both image and video benchmarks.

Similarly, VITO [Aubret et al., 2023] advocates for using videos as a natural source of supervision by identifying the most stable and discriminative elements across time. The framework aligns with human learning mechanisms by capturing dynamic scene evolution, thereby producing representations that are not only task-general but also robust to natural distribution shifts. VITO departs from prior works by questioning the efficacy of current video datasets and introducing VideoNet, a curated alternative aligned with Im-

ageNet's class distribution. This adjustment leads to enhanced spatial understanding, showing that temporally grounded training can produce features superior to those obtained through static image pretraining or adversarial methods. However, like ViC-MAE, VITO's architecture necessitates temporal sampling strategies and extensive contrastive optimization, which may restrict its application in scenarios with limited computational resources.

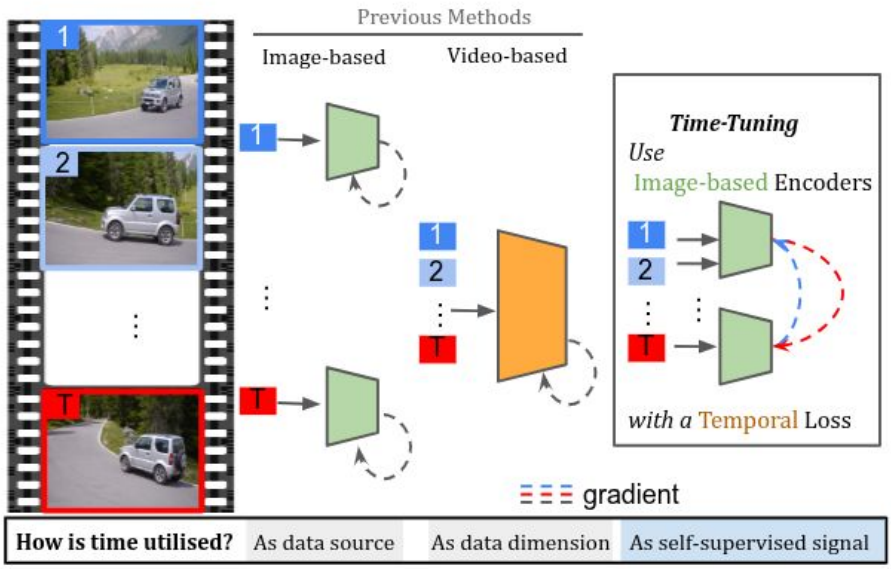

Figure 2.2: Overview of the pipeline and key discussion from the *Time Does Tell* paper [Salehi et al., 2023], which highlights how previous methods fail to fully leverage the temporal relationships between video frames. By explicitly modeling temporal consistency as a self-supervised learning signal, the proposed approach demonstrates significant performance gains. The framework shows that incorporating temporal coherence across frames leads to more robust and semantically consistent dense visual representations, benefiting both video and image-based tasks.

Source: Salehi et al. [2023]

An important advancement in this area is introduced by *Time Does Tell* [Salehi et al., 2023], which proposes an explicit and scalable framework for incorporating temporal consistency into dense self-supervised learning. Unlike previous approaches that rely on expensive 3D architectures or treat time as a mere source of augmentations, this method—illustrated in Figure 2.2—formulates a temporal self-supervised loss to fine-tune a 2D encoder initially trained on static images. By using unlabeled videos and modeling the temporal dimension as an explicit learning signal, the method achieves significant improvements not only in video-level dense prediction tasks but also in image-level semantic segmentation, showing the bidirectional transferability of video knowledge.

The core contribution of TIMET lies in its novel temporal training strategy, re-

ferred to as *time-tuning*, which leverages all frames in a sequence to construct temporally consistent dense representations. This approach goes beyond sparse frame sampling by fully exploiting the rich information distributed across time. Two main modules are introduced to achieve this: the *Feature Forwarder* (FF), which resolves the chicken-and-egg problem of establishing correspondences without ground truth labels, and a spatio-temporal dense clustering module that enforces semantic consistency across space and time. Together, these components enable the model to learn stable pixel-level features across frames without requiring supervision or synthetic correspondences.

However, while TIMET successfully demonstrates the ability to transfer temporal coherence from videos to images, its design remains closely tailored to dense prediction tasks such as unsupervised semantic segmentation. The reliance on pixel-wise consistency and spatio-temporal clustering makes it less straightforward to apply to image-level classification or global embedding tasks. Moreover, the dense computation over all video frames, though effective, introduces considerable computational overhead, which may limit its applicability in scenarios that demand lightweight inference or adaptation with restricted resources.

More broadly, a key limitation shared by several video-based methods—including TIMET and others—is their dependence on dense pixel-level supervision signals, which do not necessarily align with objectives like image retrieval, global representation learning, or lightweight adaptation. Furthermore, these methods often require access to large-scale curated video datasets and involve careful tuning for each downstream task, which can hinder the generality and plug-and-play reusability of the learned representations. Thus, while time-aware dense learning frameworks push the boundaries of what can be learned from unlabeled videos, they also underscore the need for more general, scalable, and adaptable strategies for video-to-image knowledge transfer.

In contrast, our approach focuses on learning from short, object-centric videos characterized by minimal background variation and semantic focus. We propose a lightweight adaptation strategy that avoids complex frame selection heuristics and architectural overhauls. By leveraging natural object motion across frames, we guide the model to learn invariant and generalizable representations using only a self-supervised objective. Our method stands out by requiring no labeled data and no fine-tuning, making it especially suitable for scenarios with limited supervision or restricted compute.

Crucially, our approach treats the video not as a dense temporal signal to be reconstructed, but as a rich set of diverse but semantically consistent views. This allows us to extract generalizable features from object appearance variation, occlusion, and scale, which are often missed when training on single images. By framing the problem as representation learning across intra-object viewpoints, we show that even low-resolution, clutter-free videos can enable strong transfer to image-level classification tasks.

Overall, our work advocates for a shift in how video information is used in self-

supervised learning — not as a high-fidelity temporal stream to be fully modeled, but as a simple and effective source of structured visual diversity. This enables practical, scalable, and domain-adaptive knowledge transfer from video to image tasks.

Chapter 3

Background

Our method builds on recent advances in self-supervised learning and parameter-efficient adaptation. We focus on two core components: DINO [Caron et al., 2021], a self-distillation framework for label-free representation learning, and LoRA [Hu et al., 2021], a lightweight technique for adapting large models with minimal trainable parameters. We review them in the following.

3.1 DINO: Self-Distillation with No Labels

DINO [Caron et al., 2021] is a self-supervised learning framework designed to learn powerful and transferable visual representations without the need for human annotations. It builds upon the concept of knowledge distillation, where a *student* network is trained to match the output distributions of a *teacher* network. Unlike standard distillation settings, both networks in DINO are initialized from scratch and operate on different augmented views of the same image. This architecture promotes the learning of semantically meaningful and view-invariant representations.

In DINO, the multi-crop strategy plays a crucial role in enforcing scale-invariant and semantically consistent representations. This strategy generates multiple augmented views of the same image, categorized into two types: *global crops* and *local crops*. Global crops are large image patches (e.g., 224×224 pixels) that capture most of the object and its surrounding context, enabling the model to learn high-level semantic features that are robust to significant spatial transformations. Local crops, in contrast, are smaller patches (e.g., 96×96 pixels) that contain only a portion of the object, encouraging the model to focus on fine-grained details and local patterns. By requiring consistent embeddings across both global and local crops, DINO enforces strong invariance to scale changes and viewpoint variations.

A central innovation in DINO is the asymmetric optimization of the student and teacher networks. While the student network is updated through standard gradient-

based backpropagation, the teacher network is updated using an exponential moving average (EMA) of the student's weights. This stabilizes training by introducing temporal smoothing and avoids representation collapse. The EMA update rule is defined as:

$$\theta_t \leftarrow \tau\theta_t + (1-\tau)\theta_s, \tag{3.1}$$

where θ_t and θ_s denote the parameters of the teacher and student networks, respectively, and $\tau \in [0, 1)$ is a momentum coefficient that governs the update rate.

Figure 3.1 illustrates the core components and data flow of the DINO architecture. A single input image x is transformed into two distinct augmented views, x_1 and x_2, using a diverse set of stochastic augmentations, such as cropping, color jittering, and Gaussian blur. These views are then passed through two different encoders: the student encoder g_s, which processes both local and global crops, and the teacher encoder g_t, which only receives global crops. Each encoder is followed by a projection head that outputs a feature vector, which is then normalized and passed through a softmax function to obtain a probability distribution over a fixed number of dimensions.

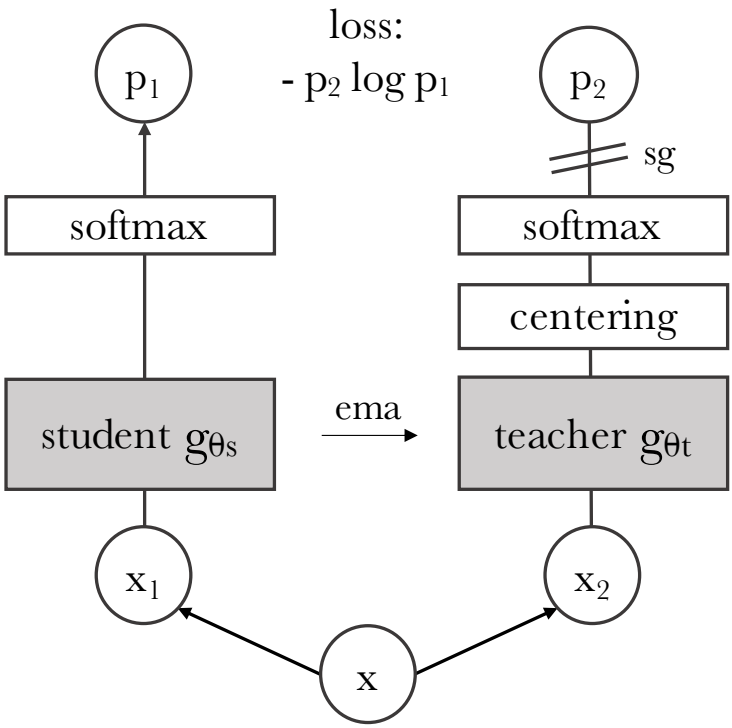

Figure 3.1: Illustration of the DINO framework adapted from Caron et al. [2021]. An input image x is transformed into multiple views. The student network is trained to predict the output distribution of the teacher network, which is updated via exponential moving average (EMA) of the student's parameters.

Source: Caron et al. [2021]

This multi-crop strategy, in which the student receives more diverse input views

than the teacher, compels the model to align representations across different spatial resolutions and contexts, thus fostering the emergence of object-centric features.

The outputs of both student and teacher networks are transformed into probability distributions using a temperature-scaled softmax function. Given an augmented view $x_{s,i}$ for the student and $x_{t,i}$ for the teacher, the output distributions are computed as:

$$f_s(x_{s,i}) = \text{softmax}\left(\frac{g_s(x_{s,i})}{T_s}\right), \quad f_t(x_{t,i}) = \text{softmax}\left(\frac{g_t(x_{t,i})}{T_t}\right), \tag{3.2}$$

where $g_s(\cdot)$ and $g_t(\cdot)$ are the respective projection outputs of the student and teacher networks, and T_s, T_t are temperature parameters. Typically, the teacher temperature T_t is lower than the student temperature T_s, resulting in sharper and more informative target distributions for the student to match.

To further prevent representation collapse, DINO employs a centering mechanism on the teacher outputs. This mechanism subtracts a running mean c from the teacher's projection before applying the softmax operation, yielding:

$$f_t(x_{t,i}) = \text{softmax}\left(\frac{g_t(x_{t,i}) - c}{T_t}\right), \tag{3.3}$$

where the center c is updated at each training step using a momentum-based moving average: $c \leftarrow \lambda c + (1-\lambda)\text{mean}_i(g_t(x_{t,i}))$, with momentum coefficient $\lambda \in [0, 1)$.

The final training objective of DINO aligns the teacher and student distributions by minimizing the cross-entropy loss across all image views:

$$\mathcal{L}_{\text{DINO}} = -\sum_i f_t(x_{t,i}) \log f_s(x_{s,i}), \tag{3.4}$$

This loss encourages the student network to produce representations that are consistent with those of the teacher, across different views of the same image. As training progresses, the student gradually acquires semantically meaningful features, which are then propagated to the teacher via the EMA update. This self-reinforcing mechanism has proven highly effective, enabling DINO to learn features that are robust, transferable, and well-suited for a wide range of downstream visual tasks.

In summary, DINO offers a powerful and scalable framework for self-supervised visual representation learning. Its architectural design and optimization strategy enable the emergence of structured and generalizable features without requiring human annotations. In the context of our work, DINO serves as the backbone for representation learning, upon which we introduce task-specific adaptations. In the following section, we examine how such adaptations can be efficiently incorporated using low-rank techniques.

3.2 Low-Rank Adaptation (LoRA)

Low-Rank Adaptation (LoRA) [Hu et al., 2021] is a parameter-efficient fine-tuning technique developed to adapt large pre-trained models with significantly fewer trainable parameters. The key idea behind LoRA is to approximate the updates to weight matrices using a low-rank decomposition, thereby avoiding the need to fine-tune the entire parameter space of the model. This approach is particularly useful when deploying large models in resource-constrained settings, such as edge devices or scenarios requiring fast adaptation with limited compute.

Consider a linear transformation in a neural network layer represented by a weight matrix $W \in \mathbb{R}^{d\times k}$. Standard fine-tuning strategies update the entire matrix W, which is computationally expensive for large models. In contrast, LoRA keeps the original weights W frozen and introduces a learnable low-rank perturbation ΔW in the form:

$$\Delta W = AB, \quad A \in \mathbb{R}^{d\times r}, \quad B \in \mathbb{R}^{r\times k}, \quad r \ll \min(d, k), \tag{3.5}$$

where A and B are small trainable matrices that together approximate the weight update. The effective weight matrix used during the forward pass becomes:

$$W' = W + \Delta W = W + AB. \tag{3.6}$$

This formulation allows the optimization process to focus only on a low-dimensional subspace of the full parameter space, significantly reducing memory consumption and training cost, while still achieving competitive performance. The rank r acts as a tunable hyperparameter that controls the trade-off between expressiveness and efficiency.

Figure 3.2 provides a schematic overview of the LoRA mechanism. During training, only the matrices A and B are updated, while the base model parameters W remain unchanged. The outputs from the low-rank adaptation are then aggregated with the frozen pre-trained weights to form the final linear transformation. This modular design enables LoRA to be seamlessly integrated into various model architectures, including transformers and convolutional networks.

A critical design feature of LoRA is its compatibility with pre-trained models. Since the base weights remain untouched, LoRA can be applied non-destructively: the adaptation can be removed at inference time or composed with other tasks. Additionally, this structure facilitates the sharing of base models across multiple downstream tasks by storing only lightweight adaptation modules.

In transformer architectures, LoRA is typically applied to the query and value projection matrices of the attention mechanism. Let $x \in \mathbb{R}^{n\times d}$ be the input sequence, and

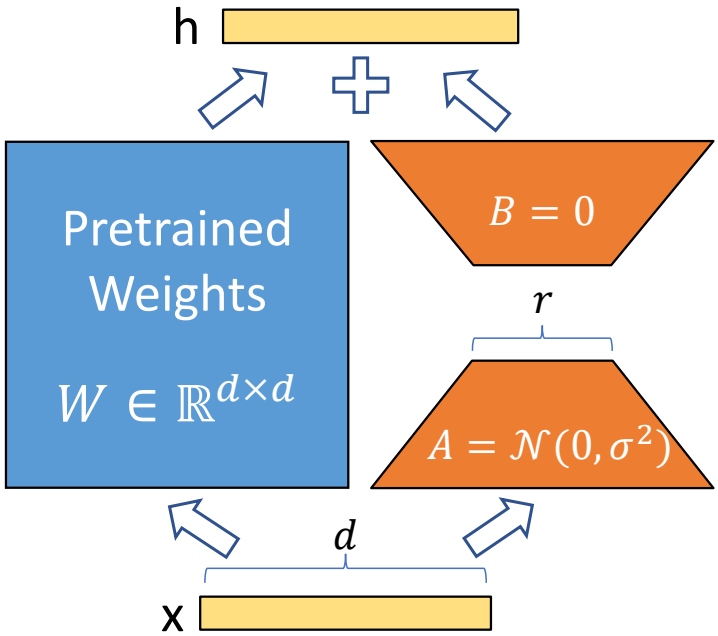

Figure 3.2: Illustration of Low-Rank Adaptation (LoRA), adapted from [Hu et al., 2021]. Instead of fine-tuning the full weight matrix W, LoRA introduces a low-rank update $\Delta W = AB$, where only A and B are trained. The update is then aggregated with the frozen base weights during the forward pass.

Source: Hu et al. [2021]

$W_q \in \mathbb{R}^{d \times k}$ the frozen query projection matrix. With LoRA, the query transformation becomes:

$$xW_q \to x(W_q + A_q B_q), \tag{3.7}$$

where $A_q \in \mathbb{R}^{d \times r}$, $B_q \in \mathbb{R}^{r \times k}$, and only A_q, B_q are trainable. This modification retains the inductive bias and pre-trained knowledge encoded in W_q, while allowing for efficient task-specific adaptation through the low-rank matrices.

To stabilize training and initialization, LoRA often scales the output of AB by a factor α/r, where α is a user-defined scaling factor. The final form of the update becomes:

$$\Delta W = \frac{\alpha}{r} AB. \tag{3.8}$$

This scaling ensures that the initial magnitude of the low-rank updates does not overwhelm the frozen weights, which is especially important when starting from random initialization of A and B.

LoRA has demonstrated strong empirical performance across a variety of domains, including natural language processing, computer vision, and multi-modal tasks. Its efficiency, modularity, and adaptability make it an appealing choice for fine-tuning large foundation models in practice.

Taken together, LoRA and DINO provide a complementary foundation for our research: while DINO enables the extraction of rich and versatile features from unlabeled visual data, LoRA facilitates efficient adaptation to new domains or tasks without the need for retraining the entire model. In the next chapter, we present our proposed methodology, which integrates these two components into a unified framework tailored to our research problem.

Chapter 4

Methodology: VESSA

This chapter presents the methodological foundation of this work, introducing VESSA (*Video-based Efficient Self-Supervised Adaptation for foundation models*) in full detail. We describe the key components, design decisions, and training strategies that compose the proposed approach. The objective is to provide a comprehensive understanding of how VESSA adapts vision foundation models to new domains using unlabeled video data. This addresses the critical challenge of domain shift in scenarios lacking annotated data. Each aspect of the method is carefully explained to highlight its contribution to the overall adaptation process.

4.1 Video-based Efficient Self-Supervised Adaptation (VESSA)

The VESSA pipeline is illustrated in Figure 4.1, and consists of three main modules: *Frame Selection*, *Preprocessing and Augmentation*, and *Model Fine-tuning*. The first module, *Frame Selection*, is responsible for selecting frames from each video and organizing them into pairs. These frame pairs are chosen to preserve object identity while capturing variations in viewpoint, providing the basis for learning view-invariant representations. The second module, *Preprocessing and Augmentation*, performs standard preprocessing steps and applies stochastic transformations to each frame in the pair independently. These transformations increase input diversity and support robust representation learning by encouraging invariance to appearance-level changes. The third module, *Model Fine-tuning*, implements the adaptation strategy. It updates the weights of a pretrained vision foundation model using a self-supervised learning objective that exploits the relationships between the paired augmented views. This step enables domain adaptation without requiring labeled data. Each of these modules will be described in detail in the following subsections.

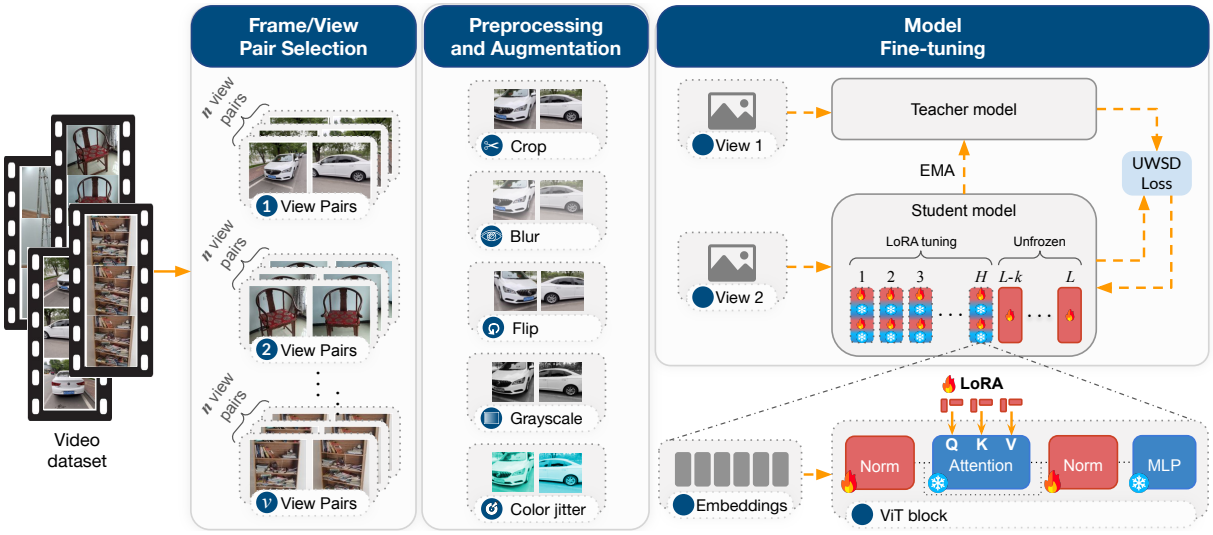

Figure 4.1: **The proposed training pipeline.** The model input consists of videos, which first undergo a *Frame Selection* stage, where n pairs of frames are sampled from each video. These pairs are then passed through the *Preprocessing and Augmentation* stage, where distinct transformations are applied to the first and second images of each pair. The resulting views are fed into teacher and student networks, both initialized from the same foundation model; LoRA is applied to their architectures for parameter-efficient fine-tuning. Finally, the *uncertainty-weighted self-distillation loss* (UWSD) is applied to align their representations.

4.1.1 Frame Selection

The first stage of our adaptation pipeline focuses on extracting informative and diverse visual signals from unlabeled video data. Since videos naturally contain object-centric sequences with temporal coherence, the objective of this module is to construct frame pairs that capture viewpoint variations while maintaining semantic consistency. This stage enables a novel use of image-based VFMs by allowing them to extract supervision signals from video sequences, bridging the gap between static-image pretraining and dynamic visual inputs.

Each input video V, with T frames $\{F_i\}_{i=1}^{T}$ is first processed by the Frame Selection module, which samples n frame pairs per video. For each pair of frames that we aim to construct, we first randomly sample a starting frame index $t \sim \mathcal{U}(1, T - \delta_{\max})$, where $\delta_{\max}$ is a predefined maximum temporal offset, and $\mathcal{U}$ denotes the uniform distribution. Then, we sample frames based on a temporal gap $\delta \sim \mathcal{U}(1, \delta_{\max})$, ensuring a minimum offset of one frame. Each selected pair of frames is then formed according to the rule:

$$\delta \in [1, \delta_{\max}], \quad t \in [1, T - \delta]$$

This randomized strategy introduces temporal diversity by allowing variable dis-

tances between frames, which helps the model learn more robust representations across different viewpoints. At the end of this module, we obtain a batch composed of frame pairs sampled from different videos, represented as:

$$\mathcal{B} = \{(F_{t_k}, F_{t_k+\delta_k}) \mid k = 1, \ldots, v\},$$

where each $(F_{t_k}, F_{t_k+\delta_k})$ is a pair of frames from the k-th video in the batch, and v is the batch size. The temporal index t_k and offset δ_k are sampled independently for each pair. This setup ensures that the batch contains diverse video content and temporal gaps.

By sampling frame pairs with variable temporal gaps across different videos, this module creates a rich training batch with both spatial and temporal diversity. The resulting collection of frame pairs forms the foundation for the subsequent preprocessing and augmentation steps, where appearance variations are introduced to further enhance representation robustness.

4.1.2 Preprocessing and Augmentation

Once the frame pairs are selected, the next step is to introduce appearance-level variability while maintaining temporal structure. The Preprocessing and Augmentation module applies distinct transformation pipelines to each frame in the pair, promoting robustness to changes in lighting, texture, and spatial composition. This is essential for building representations that are both discriminative and invariant to superficial differences.

In the Preprocessing and Augmentation module, each frame in the pair k is transformed independently using two distinct pipelines of random augmentations. These operations yield two augmented views: $F_t^{(a)}$ and $F_{t+\delta}^{(b)}$, which are designed to promote appearance diversity and representation robustness. This module also generates the *local crops* used in our adaptation. Inspired by the reference method, which employs multiple local crops per image to stabilize fine-tuning, we adapt this strategy by sampling local crops as pairs—one from each frame. This pairing preserves temporal coherence while promoting local spatial variation, contributing to more robust and transferable representations. At the end of this module, we have

$$\mathcal{B} = \left\{ \left(F_{t_k}^{(a)},\ F_{t_k+\delta_k}^{(b)},\ \left\{ F_{t_k}^{(c_i)},\ F_{t_k+\delta_k}^{(c_i)} \right\}_{i=1}^{u} \right) \mid k = 1, \ldots, v \right\},$$

where:

- $F_{t_k}^{(a)}$ and $F_{t_k+\delta_k}^{(b)}$: two temporally separated frames from the k-th video, with distinct global transformations a and b applied;

- $\left\{F_{t_k}^{(c_i)},\ F_{t_k+\delta_k}^{(c_i)}\right\}_{i=1}^{u}$: a set of u pairs of local crops, where c_i denotes a distinct transformation involving a small crop on the main frame; and u the number of local crop pairs.

Through the use of global augmentations and temporally aligned local crops, this module ensures that the input to the model contains complementary views of the same object. These diverse and structured inputs prepare the model for the fine-tuning phase, where temporal alignment and transformation invariance are leveraged as self-supervised signals for representation learning.

4.1.3 Model Fine-tuning

The final component of the VESSA pipeline is responsible for adapting the pretrained vision foundation model to the target domain using the prepared augmented frame pairs. This is achieved through a self-distillation framework, where both global and local views are processed by teacher and student networks to align their representations. To ensure efficient fine-tuning with minimal supervision, we leverage a parameter-efficient strategy based on LoRA and introduce an uncertainty-aware loss formulation.

In the Model Fine-tuning stage, training is performed in batches; however, for clarity, we describe the process using a single pair of images and associated local crops. The pair k, given by $(F_t^{(a)}, F_{t+\delta}^{(b)})$, is passed through both the student and teacher networks, while the local crops are processed only by the student. Both networks are initialized from the same pretrained vision foundation model. The outputs of the student and teacher are denoted as follows:

$$s = f_{\mathrm{s}}(F_t^{(a)}), \quad q = f_{\mathrm{t}}(F_{t+\delta}^{(b)}), \quad s_{lc1} = f_{\mathrm{s}}(F_t^{(c)}), \quad s_{lc2} = f_{\mathrm{s}}(F_{t+\delta}^{(c)})$$

s, q denote the outputs of the projection head applied to the global frame representations. The final representations of all local crops (i.e., s_{lc1} and s_{lc2}), along with s, are also compared to q as part of the DINO loss computation. The fine-tuning training objective is a weighted form of the formulation presented in Chapter 3. To prioritize uncertain teacher outputs, we introduce an *Uncertainty-Weighted Self-Distillation (UWSD)* loss, which modulates the contribution of each sample to the loss based on the estimated uncertainty of the teacher's predictions. We compute the entropy of the teacher's output distribution, which then informs the uncertainty-based weighting function:

$$w(q) = 1 + \gamma \cdot \mathcal{H}(q),$$

where γ is a hyperparameter controlling the influence of uncertainty. The final training objective becomes:

$$\mathcal{L}_{\text{UWSD}} = \frac{1}{N} \sum_{(q,s,s_{lc_i}) \in \mathcal{B}} w(q) \cdot \mathcal{L}_{\text{DINO}}(q, s, s_{lc_i})$$

This fine-tuning step completes the adaptation pipeline by aligning the student's representations with the teacher's under the guidance of uncertainty-weighted supervision. The combination of global and local losses, parameter-efficient updates, and temporal supervision allows VESSA to effectively adapt to new domains using only unlabeled video data. This stage sets the groundwork for the optimization refinements discussed in the next section.

4.2 Critical optimization considerations

Adapting pretrained foundation models to new domains via self-supervised fine-tuning introduces a number of optimization challenges. These arise primarily from distributional shifts between pretraining and target datasets, as well as from architectural inconsistencies introduced by randomly initialized projection heads. To address these issues, this section presents a set of training strategies designed to ensure a stable and effective continuation of self-supervised learning.

Continuing self-supervised training from a pretrained VFM poses significant challenges, particularly due to gradient instabilities that emerge when shifting to a new domain. These instabilities are often caused by the discrepancy between the pretrained distribution and the target data, as well as by the abrupt introduction of a randomly initialized projection head.

Foundation models such as DINO are pretrained over hundreds of epochs with a joint optimization of both the backbone and the projection head, resulting in a well-aligned embedding space. When fine-tuning on a new dataset—often smaller and semantically distant from the pretraining corpus—the backbone is typically retained, but a new projection head is initialized. This mismatch leads to sharp changes in gradient flow, causing the model to rapidly forget useful representations and struggle with optimization convergence.

In standard training regimes, all network parameters are typically updated simultaneously, which can lead to unbalanced gradient flow and further degrade the pretrained representations. To mitigate this, we initially freeze the backbone and train only the projection head for a few epochs, allowing it to adapt to the existing embedding space. As

another strategy to alleviate representational drift, we gradually unfreeze the backbone, applying different update strategies across its layers.

Specifically, we enable fine-tuning of the first H layers using LoRA [Hu et al., 2021], which restricts updates to low-rank adaptations of the attention weights—namely in the Query, Key, and Value projections of each self-attention layer—while keeping normalization layers trainable. This design helps preserve low-level visual features such as edges and textures, which are broadly transferable across domains. In contrast, the last L layers of the backbone are fully unfrozen and updated using standard gradient descent, allowing high-level semantic features to adapt to the new data. This staged unfreezing strategy ensures training stability, maintains prior knowledge, and enhances adaptation efficiency.

By gradually unfreezing the backbone, applying LoRA to lower layers, and preserving key normalization parameters, our approach maintains the generalization capabilities of the pretrained model while enabling domain-specific adaptation. These strategies collectively improve training stability, reduce representational drift, and enhance the final performance of the adapted model. The effectiveness of these techniques will be validated in the experimental results that follow.

After presenting the methodology, we now provide an overview of the complete training pipeline. The videos used in our experiments are short, object-centric, and have a standardized background, as illustrated in Fig. 5.2. The pipeline of VESSA follows five main steps. First, the video dataset is preprocessed by extracting frame sequences from each short video and generating positive training pairs from distinct frames of the same sequence. Second, the pretrained backbone is loaded, and a LoRA configuration is applied to selected transformer layers, followed by the initialization of the projection head. Third, in the first training stage, the backbone parameters are frozen, and only the projection head is optimized for n_1 epochs using the self-distillation loss computed between student and teacher embeddings of the paired frames. Fourth, during the second training stage, backbone layers are progressively unfrozen according to a predefined schedule, enabling the joint optimization of the projection head and LoRA parameters while keeping the remaining layers frozen. Finally, the adapted model is evaluated on downstream classification tasks to assess its performance in the target domain.

Chapter 5

Experiments

5.1 Experimental Setup

This section presents the experimental setup, including the datasets used, the training and evaluation protocols, and the statistical analyses conducted. These components support the empirical evaluation of our method and its comparison with relevant baselines. We describe the configuration of the data, model evaluation strategies, and the metrics employed to assess performance. Each subsection provides a detailed overview of a specific aspect of the experimental pipeline.

5.1.1 Datasets and protocol

The MVImageNet dataset [Yu et al., 2023] is a large-scale multi-view image dataset specifically curated to facilitate view-invariant visual representation learning. It comprises approximately 6.5 million frames extracted from over 219,000 short videos, spanning 238 object categories. Each video captures a single object instance from multiple viewpoints, typically recorded as short clips where the camera orbits around the object, enabling the capture of diverse visual cues while preserving object identity. The dataset features a wide variety of everyday object classes, including animals, vehicles, tools, and household items, offering substantial intra-class and inter-class diversity. All videos are collected under real-world conditions, reflecting natural variations in lighting, background, and occlusion. MVImageNet is hierarchically structured, with each category containing a varying number of video clips. Figure 5.3 presents the distribution of video samples per class, illustrating the inherent imbalance across categories. Figure 5.1 provides representative examples of video sequences from different object classes. For a complete list of all 238 categories included in the dataset, please refer to Appendix A.1.

The CO3D dataset [Reizenstein et al., 2021] is a large-scale dataset of object-centric videos designed to support research in 3D reconstruction, novel view synthesis, and multi-view visual learning. It comprises approximately 1.5 million frames extracted from over 19,000 real-world video clips, covering 50 common object categories such as backpacks, chairs, and cars. Each video sequence captures a single object instance from multiple viewpoints, typically recorded by a handheld device in unconstrained environments. The dataset is characterized by significant variation in lighting, occlusion, and background, making it suitable for robust representation learning under realistic conditions. Figure 5.4 shows the distribution of video sequences per category, highlighting the class imbalance inherent in the dataset. The figure 5.2 presents visual examples of object-centric videos across different categories. For a complete list of all object categories included in CO3D, please refer to Appendix A.2.

Using the aforementioned datasets, MVImageNet and CO3D, we defined a unified training and evaluation protocol. Our goal is to train on videos and evaluate on individual images. Initially, we designed a protocol that splits each class into training and testing sets using a 75%-25% ratio. For training, we construct frame pairs from the video clips following our proposed methodology. For evaluation, we sample one frame from each training video and one frame from each testing video; each frame inherits the label of its corresponding video. Our validation pipeline consists of extracting image embeddings from both training and testing samples using the vision foundation model under evaluation. Specifically, we extract the embedding corresponding to the classification token (`[CLS]`), which is the learnable token prepended to the input sequence and designed to aggregate global information. We then apply the k-Nearest Neighbors (KNN) algorithm to assign a class label to each test sample based on its proximity in the embedding space. By default, we report results with $k = 1$, meaning that each test image is assigned the label of the nearest training image in the learned representation space.

5.1.2 Implementation details

Our experiments were performed on TPU v3-8, featuring 8 cores and 128 GB of high-bandwidth memory. All implementations used the `scenic` library [Dehghani et al., 2022] in JAX. We adopted the ViT-Base and ViT-Small architectures to balance performance and efficiency, as preliminary tests showed that the tiny model underfit and the Large model offered minimal gains at higher cost. Our training followed the base hyper-parameter configuration of the DINO protocol [Caron et al., 2021], except for the specific settings detailed below. As a reference, we adopted 10 training epochs for both the initial

Figure 5.1: Representative examples of object-centric videos in MVImageNet. Each column illustrates a sequence of frames showing the same object captured from different viewpoints.

Source: Yu et al. [2023]

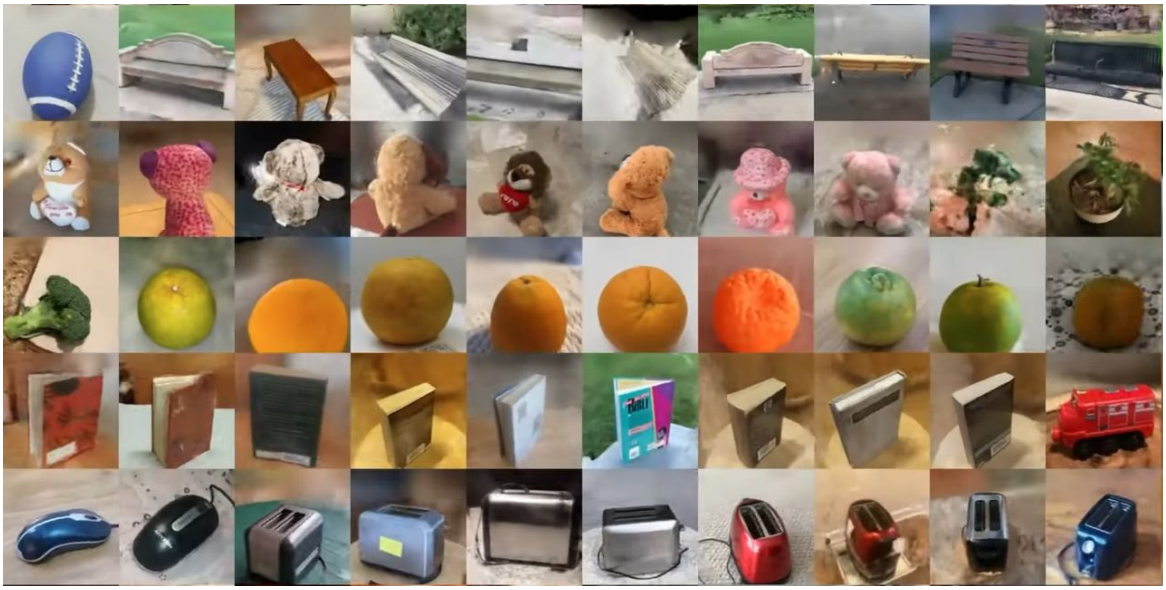

Figure 5.2: Visual examples of videos in CO3D. Each image represents a single frame sampled from an object-centric video, captured from a specific viewpoint under real-world conditions. Different categories and viewpoints are illustrated across the samples.

Source: Reizenstein et al. [2021]

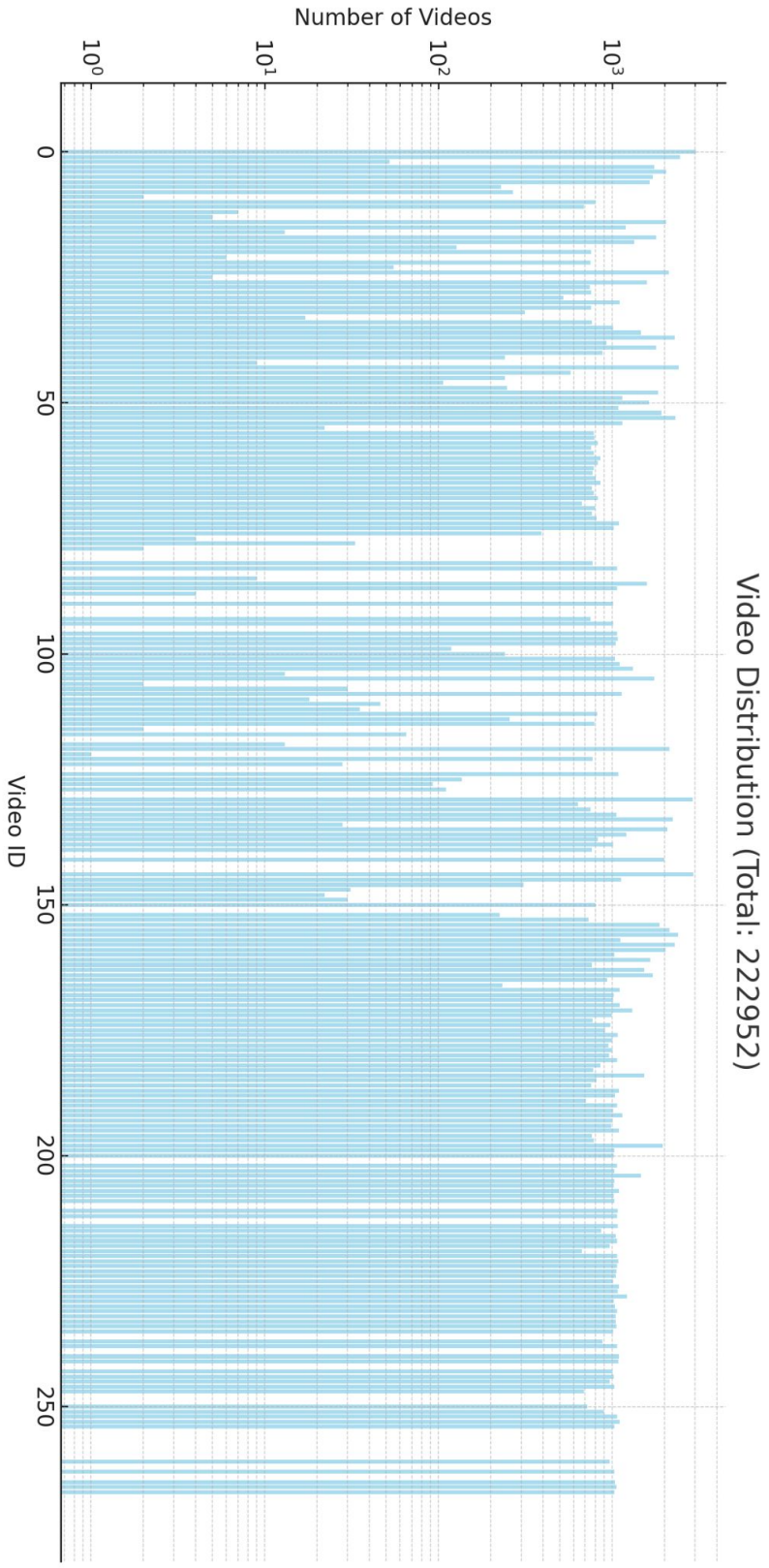

Figure 5.3: Distribution of the number of video clips per class in the MVImageNet dataset. The dataset presents a natural class imbalance, with some object categories being significantly more represented than others.

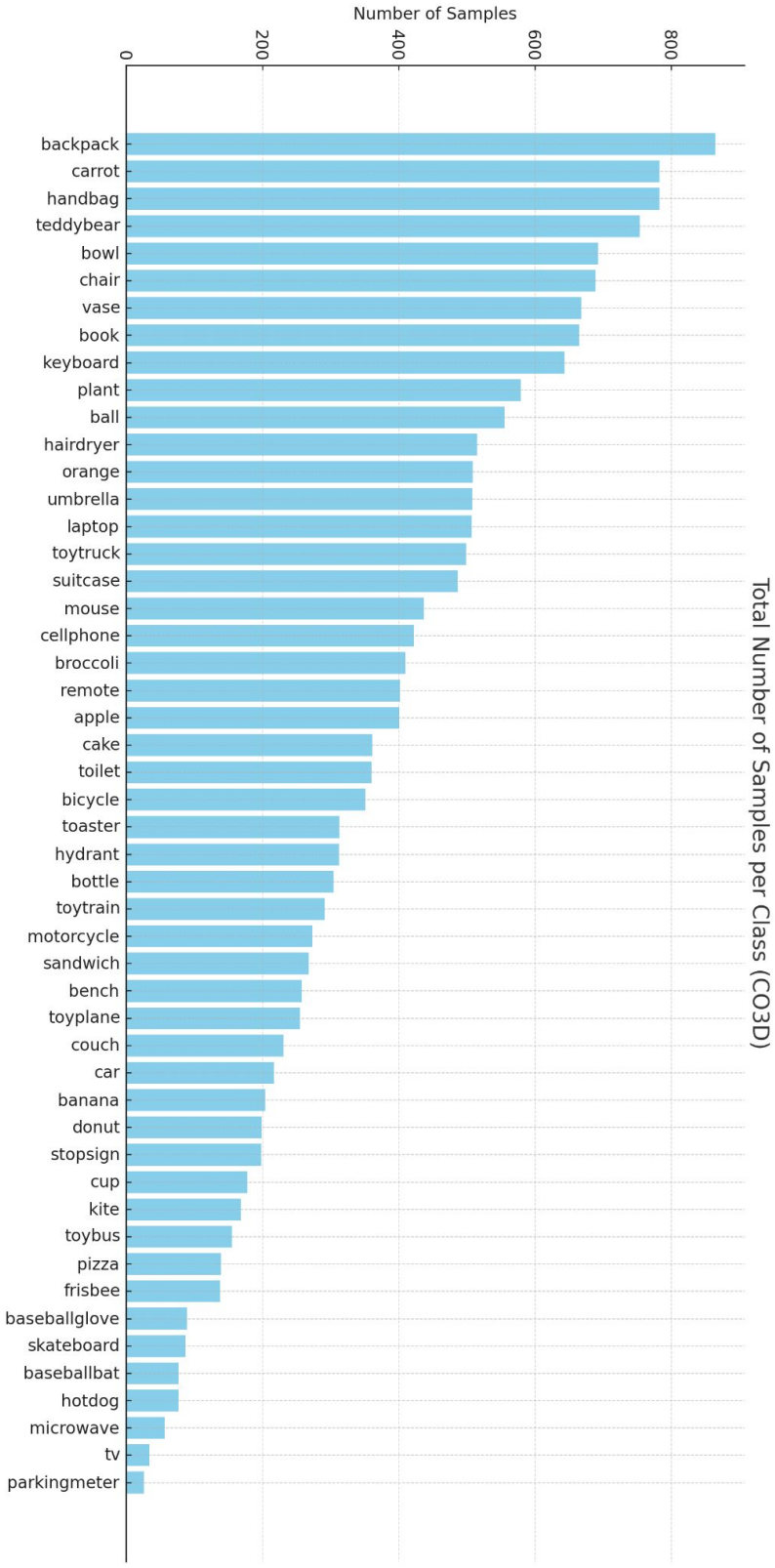

Figure 5.4: Distribution of the number of video sequences per category in the CO3D dataset. The dataset exhibits significant variation in the quantity of samples across different object classes.

projection head adaptation and the subsequent full model training, using a batch size of 256 and an input image resolution of 224×224. For each video, we sampled 3 frame pairs. The hyperparameter γ, which controls the weight of the distillation loss, was set to 1. For the image-based baseline, we used the first frame of each pair as the reference image. This frame was then processed through the subsequent steps as a single-image input, following the standard image-based pipeline.

5.1.3 Statistical significance test

To assess the significance of the observed differences between the top-performing configurations, we conducted statistical tests comparing the two best results. Each experimental setup was run independently three times. We employed an unpaired Student's t-test with a 90% confidence level to evaluate whether the performance improvements were statistically meaningful. Confidence intervals for the main comparisons are reported in the supplementary material, highlighting cases where the differences were statistically significant.

5.1.4 Visual Foundation Models

Our experiments were conducted using widely adopted and state-of-the-art VFMs, including DINO [Caron et al., 2021], DINOv2 [Oquab et al., 2023], and TIPS [Maninis et al., 2025]. These models represent strong visual backbones commonly used in vision tasks, particularly within SSL frameworks. DINO and DINOv2, both pretrained using SSL objectives, have been discussed extensively in previous chapters. In contrast, TIPS was not trained using an SSL strategy. We include TIPS in our evaluation to highlight the generality of our proposed adaptation method, which can be applied to VFMs regardless of their original pretraining approach. Our unified protocol enables adaptation without additional pretraining constraints, and its application to TIPS further demonstrates the flexibility and effectiveness of our approach. The TIPS model is introduced in detail.

TIPS [Maninis et al., 2025] (Text-Image Pretraining with Spatial Awareness) is a multimodal foundation model that aims to improve visual-language alignment by incorporating spatial structure into the pretraining process. Unlike prior works that treat vision-language pretraining as a flat alignment between global image and text representa-

tions, TIPS introduces an auxiliary spatial-language alignment objective. This objective guides the model to associate localized image regions with semantically related phrases, improving the spatial grounding of visual features. The architecture of TIPS consists of a vision encoder and a text encoder trained jointly using both image-text contrastive loss and a spatial contrastive loss. As a result, TIPS achieves enhanced performance on downstream tasks requiring fine-grained spatial understanding, such as referring expression comprehension and visual grounding. In our study, we focus exclusively on the VFM from TIPS to evaluate the adaptability of our proposed protocol across models with diverse pretraining strategies. The authors of TIPS release both the visual and textual encoders of the model; however, in our experiments, we focus solely on the VFM to evaluate its compatibility with our adaptation protocol. TIPS demonstrates strong performance on weakly supervised recognition tasks, and its inclusion in our study further illustrates the flexibility of our method across diverse pretraining strategies.

We compare our method with ExPLoRA [Khanna et al., 2024], a recent approach for improving transfer learning of pretrained vision transformers (ViTs) under domain shifts. To ensure a fair comparison with our video-based approach, we extend ExPLoRA to operate on short object-centric videos. Experiments using TIPS + ExPLoRA were not reported once the original work ExPLoRA clearly specifies that it is designed for continual self-supervised learning using training heads from self-supervised models. TIPS, however, is not a self-supervised model.

5.2 Results

We begin our evaluation by analyzing the individual contributions of each component of the VESSA approach through a comprehensive ablation study. This analysis provides insights into the effectiveness of leveraging video data for self-supervised adaptation and highlights the critical design choices that enable our approach to outperform existing alternatives. In what follows, we systematically isolate and compare different configurations, followed by comparisons to state-of-the-art baselines across multiple datasets and model architectures.

A series of experiments conducted on the MVImageNet dataset using the vision transformer-small (ViT-S) to isolate and evaluate the contributions of each component of our method are shown in Table 5.1. Given the challenge of adapting to a new domain with limited and unlabeled data, we first trained DINO from scratch using both image and video data. As expected, the results were suboptimal due to the limited data, with a final accuracy of 33.86%. However, training with video consistently outperformed the

Table 5.1: **Ablation study on components for video-based self-supervised ViT fine-tuning.** All models use ViT-S/16 and are evaluated with k-NN (k=1) on the MVImageNet dataset. We analyze the impact of architectural choices, local crops, training heads, and data modalities.

Method	UWSD Loss	Unfrozen Last Layers	Local Crops	Train Head	Input	Accuracy (%)
VESSA	✓	2	✓	✓	Video	**91.87**
	✓	2		✓	Video	90.53
		2	✓	✓	Video	90.92
	✓	1	✓	✓	Video	87.14
	✓	3	✓	✓	Video	90.80
	✓	4	✓	✓	Video	90.55
	✓	2	✓		Video	80.87
	✓	2	✓	✓	Image	88.54
DINO			✓		Image	33.86
DINO			✓		Video	39.39
DINO Pretrained					Image	89.69

image-based counterpart, achieving 39.39% accuracy—an improvement of 5.53 percentage points (p.p.) aligning with our motivation to leverage temporal information—though the results remain relatively low overall. Subsequently, we applied the pretrained DINO model directly, which yielded solid performance and served as a strong baseline, achieving 89.69% accuracy. We also evaluated a naive continuation of training using only images, which similarly led to performance degradation, resulting in a slightly lower accuracy of 88.54%. Careful decision of the training head projection before fine-tuning had a significant impact, improving performance by approximately 10 p.p.. This result indicates the importance of carefully designed adaptations to achieve such improvements. Finally, we compared our full method against its individual variants, confirming that our complete approach achieves the best results, validating the importance of each design choice in the overall effectiveness of VESSA. The best-performing configuration achieved 91.87% accuracy, representing an improvement of 2.18 p.p. over the pretrained DINO baseline, with the optimal setting obtained by unfreezing the last 2 layers during adaptation.

Across all experiments, leveraging video consistently outperforms frame-based alternatives. To better understand the source of these gains—whether from motion cues or temporal continuity—we conducted an additional experiment on the CO3D dataset using DINO and DINOv2. Specifically, we evaluated the impact of frame distance (δ) during self-supervised fine-tuning, as detailed in Table 5.2. The results show that varying the temporal gap between frames affects performance, with the highest accuracy achieved when δ was randomly sampled from the range [5, 10], yielding 85.03% with DINO and 91.85% with DINOv2. This suggests that exposing the model to diverse temporal relationships between frames contributes positively to representation learning.

Table 5.2: **Top-1 accuracy (%) on the CO3D dataset using different frame distance strategies ViT-B.** We report k-Nearest Neighbors (k=1) classification accuracy for models pretrained with DINO and DINOv2. Each value of δ defines the temporal distance between frames selected from videos during self-supervised fine-tuning.

Frame Distance (δ)	**DINO**	**DINOv2**
1	85.00	91.51
2	84.73	91.52
3	84.90	91.39
4	84.27	91.42
5	84.66	91.80
10	85.00	91.74
15	84.54	91.25
20	84.82	91.23
Random [5, 10]	**85.03**	**91.85**
Random [10, 30]	82.46	91.52

After analyzing the individual components of our approach, we now turn to a broader evaluation of VESSA applied to different backbone models across two datasets. As shown in Table 5.3, all other methods significantly outperform the base pretrained models without fine-tuning (the first row of the tables), except for a few cases involving a baseline variant of our method that employs static images only, which we refer to as Static-baseline. In particular, when video data is used, unsupervised fine-tuning generally leads to superior performance across both datasets and architectures, with the exception of Explora with video and DINO in MVImgNet, where performance decreases relative to the pretrained model. The performance gap between VESSA and Static-baseline, as well as between the ExPLoRA baseline and ExPLoRA + video, confirms the effectiveness of adapting models to the target domain using unlabeled video data in both CO3D and MVImageNet datasets. In contrast, the results of the Static-baseline indicate that a naive image-based self-supervised continual learning approach is not sufficient to achieve successful fine-tuning.

Considering the CO3D dataset, applying VESSA to DINOv2 yields the best result of 91.85% ± 0.56, which is 2.21 p.p. higher than ExPLoRA + video (89.64 ± 0.47); this difference is statistically significant. On the other hand, for the MVImageNet dataset, VESSA achieved 96.01 ± 1.08 while ExPLoRA + video reached 96.15 ± 0.87; however, the difference is not statistically significant. For statistical comparisons, please refer to the supplementary tables (CO3D Table 5.4 and MVImagenet Table 5.5), which present pairwise comparisons between the two best-performing results along with their corresponding confidence intervals.

Nevertheless, when considering DINO, VESSA outperforms again the ExPLoRA

results. Moreover, our approach also performs better with TIPS against its pretrained version. It is important to present these results—alongside those in Table 5.1—as they demonstrate that simply continuing self-supervised training with domain-specific image data, while seemingly straightforward, does not yield consistent improvements. As the results show, this strategy often leads to performance degradation or, at best, no noticeable gains.

Table 5.3: **Top-1 accuracy (%) on CO3D and MVImageNet datasets using k-Nearest Neighbors (k=1).** We compare pretrained vision foundation models (Dino [Caron et al., 2021], Dinov2 [Oquab et al., 2023] and TIPS [Maninis et al., 2025]), an image-based baseline (ExPLoRA [Khanna et al., 2024]), and our proposed video-based fine-tuning method. All results are reported on the validation set using representations extracted from the backbone and evaluated via KNN. Our method achieves superior performance by leveraging object-centric videos for unsupervised adaptation.

	CO3D			MVImageNet		
Method	**DINO**	**DINOv2**	**TIPS**	**DINO**	**DINOv2**	**TIPS**
Pretrained	78.86	87.86	60.02	90.44	95.75	78.65
ExPLoRA	79.78	88.31	—	90.94	95.79	—
ExPLoRA+video	83.64	89.64	—	87.74	**96.15**	—
Static-baseline	80.31	81.60	55.59	89.39	92.53	76.05
VESSA (ours)	**85.03**	**91.85**	**70.56**	**92.51**	96.01	**80.54**

Table 5.4: **Top-1 accuracy (%) on the CO3D dataset using k-Nearest Neighbors (k=1).** We compare pretrained vision foundation models, an image-based baseline, and our proposed video-based fine-tuning method. ExPLoRA and VESSA results are reported on the validation set using representations extracted from the backbone and evaluated via k-NN. We report confidence intervals to highlight the statistical significance of the improvements.

Method	**DINO-B**	**DINOv2**	**TIPS**
ExPLoRA + video	83.64 ± 0.84	89.64 ± 0.47	—
VESSA (ours)	**85.03** ± 0.52	**91.85**± 0.56	**70.56**± 1.03

To qualitatively illustrate the benefits of video-based self-supervised training, Figure 5.2 presents examples of top-1 nearest neighbor retrievals based on the learned embeddings. When comparing the pretrained DINOv2 model with our proposed VESSA method, we observe that DINOv2 often produces embeddings influenced primarily by background and global scene features. In contrast, VESSA consistently focuses on the object of interest, even in challenging scenarios where the appearance of the retrieved object differs in color or texture from the query image. This suggests that VESSA learns

Table 5.5: **Top-1 accuracy (%) on the MVImageNet dataset using k-Nearest Neighbors (k=1).** We compare pretrained vision foundation models, an image-based baseline, and our proposed video-based fine-tuning method. ExPLoRA and VESSA results are reported on the validation set using representations extracted from the backbone and evaluated via k-NN. We report confidence intervals to highlight the statistical significance of the improvements.

Method	**DINO-B**	**DINOv2-B**	**TIPS-B**
ExPLoRA + video	87.74 ± 1.03	**96.15** ± 0.87	—
VESSA (ours)	**92.51** ± 1.11	96.01 ± 1.08	**80.54** ± 1.71

more semantically meaningful and object-centric representations, improving robustness and task relevance.

For instance, in the first column, the query image contains a ball placed on a multicolored striped background. DINOv2 retrieves an image of an orange, which shares similar shape and background characteristics but belongs to a different semantic class. VESSA, on the other hand, retrieves an image of a ball with both visual and contextual similarity, correctly attending to the object rather than the scene. This behavior generalizes across the examples. Two particularly illustrative cases are shown in the second and fifth columns. In the second column, a baseball glove is misclassified by DINOv2 as a plant—likely due to background similarities in texture and color—whereas VESSA correctly retrieves another glove. In the final column, DINOv2 confuses a carrot with a keyboard, again prioritizing background features, while VESSA accurately retrieves a carrot with a similar scene composition. These results demonstrate that VESSA produces embeddings that are better aligned with object semantics, enhancing retrieval accuracy and interpretability.

We now turn to a set of complementary experiments aimed at providing further insights into the behavior and characteristics of our proposed adaptation technique. Specifically, these analyses explore three key questions: how the input structure used in our method—based on temporally coherent video frame pairs—differs from standard image-based self-supervised approaches; whether our method is affected by catastrophic forgetting, using the original foundation model as a reference point; and whether comparable performance could be achieved by applying purely geometric transformations to single images, without relying on multi-frame video inputs. Together, these experiments offer a broader understanding of the mechanisms and limitations of video-based adaptation in vision foundation models.

As a first complementary analysis, to gain a deeper understanding of how the foundation model behaves after the adaptation process with VESSA, we conducted an experiment designed to assess its susceptibility to forgetting. Specifically, training was

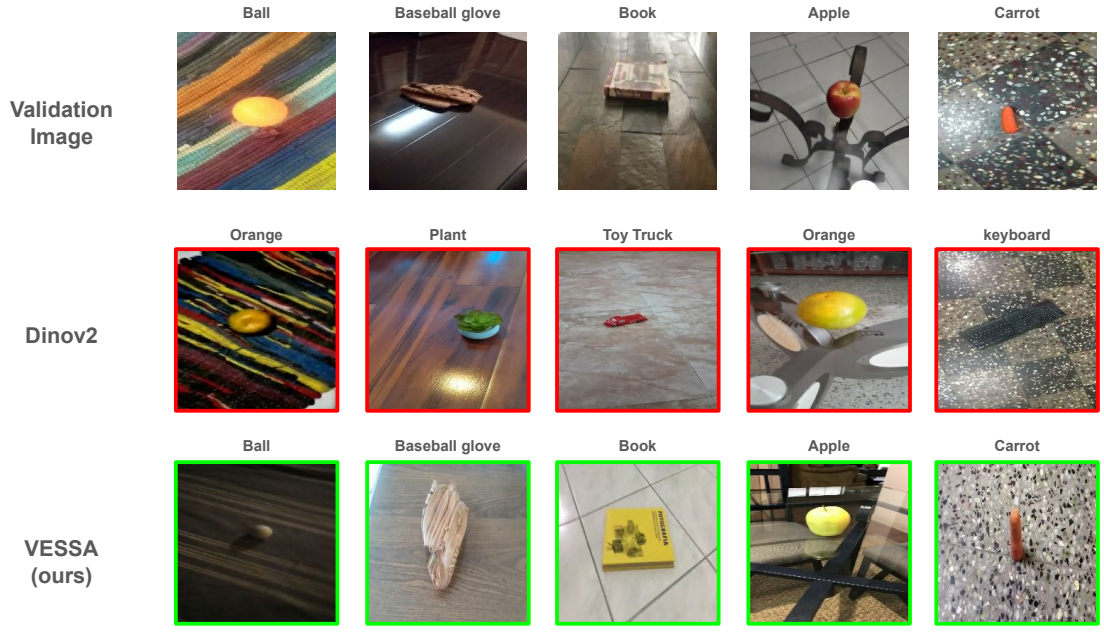

Figure 5.5: Qualitative examples of nearest neighbor retrieval ($k = 1$) on the CO3D validation set. We selected some of the most challenging validation samples and retrieved the nearest neighbors using two methods, shown row-wise: the first row displays the query (validation) images; the second row presents the retrievals using raw DINOv2 features; the third row shows the retrievals produced by our method. Images with red borders indicate incorrect retrievals, while green-bordered images represent correct ones. These examples illustrate the effectiveness of leveraging multi-frame video information for representation learning. Notably, our method demonstrates greater focus on the object of interest, whereas the baseline often retrieves matches dominated by background similarity.

performed using VESSA exclusively on the MVImageNet dataset, followed by evaluation on the held-out test set of the CO3D dataset. As shown in Table 5.6, this cross-dataset setting reveals a marked drop in performance—approximately 5 to 7 percentage points—when compared to the baseline results obtained by pretraining and evaluating on the same dataset. This performance degradation highlights both the presence of catastrophic forgetting and the limited generalization capabilities of the adapted model when exposed to a distribution shift, even when trained on a diverse and temporally rich video corpus.

As a second complementary analysis, aiming to qualitatively assess the impact of using video-based inputs—a core component of our approach—we highlight the differences between the view generation strategies employed by VESSA and those used in standard image-based methods such as DINO. To that end, we present illustrative examples of view pairs from both approaches. It is important to note that in both VESSA and DINO, the same sets of transformations are applied independently to each view. To assess the impact of video-based inputs on representation learning, we analyze input variability by

Table 5.6: Performance comparison between DINO and DINOv2 models using the pretrained base model, our proposed VESSA method, and the cross-dataset evaluation. The cross-dataset model was trained on MVImageNet and tested on CO3D to analyze forgetting behavior, demonstrating the degradation experienced when a model is trained on one dataset and evaluated on another. All experiments utilized the ViT-B architecture.

Method	**DINO-B**	**DINOv2-B**
Pretrained	78.86	87.86
Cross dataset	74.40	80.36

contrasting the frame selection strategy used in VESSA with the standard augmentation-based sampling in DINO and DINOv2 (see Appendix A.3 for full details of the applied transformations). As shown in Figure 5.6, frame pairs selected by VESSA—based on a fixed temporal offset of $\delta = 5$ frames—exhibit substantially greater visual diversity than those generated through standard augmentations. In contrast, the views generated by DINO/DINOv2 tend to be more visually homogeneous. This enhanced variability introduced by real video frames is likely a key factor in the performance differences observed when training with video data, as opposed to relying solely on static image augmentations.

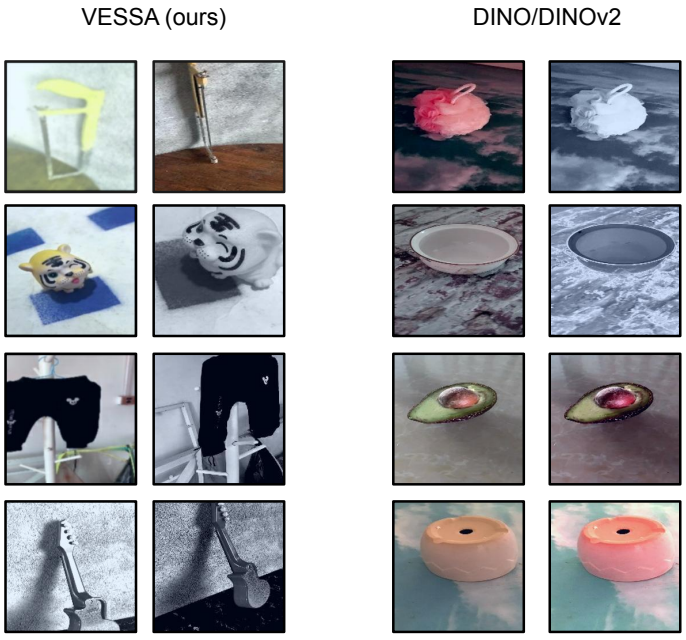

Figure 5.6: Example frames from the MVImageNet dataset illustrating the differences between global crop input pairs used for the teacher and student networks during training with DINO and VESSA (ours). Our method, VESSA, introduces substantially greater variability in the appearance of the evaluated object. The temporal distance between the selected frames is $\delta = 5$ frames. The first image of each pair shows the global crop from the transformation of view 1, and the second image of each pair shows the global crop corresponding to the transformations of view 2.

Table 5.7: Performance comparison using our method and images and transformations to simulate camera movement in images with DINO and DINOv2 on the CO3D dataset with $k = 1$.

Method	**DINO - B**	**DINOv2 - B**
VESSA	**85.03**	**91.85**
Static-baseline	80.31	81.60
Static-baseline + Transf. simulate video	80.60	81.49

As the third and final complementary experiment, motivated by the strong performance observed when training with videos, we investigated whether additional image transformations—beyond those used in the standard DINO pipeline—could simulate the benefits of camera motion. To this end, we applied a set of motion-inspired augmentations to one of the views during training, aiming to mimic the effect of slight viewpoint changes. Specifically, we incorporated translations of up to 10% of the image dimensions, rotations up to 10 degrees, scaling variations up to 5%, brightness shifts of 0.1, and contrast adjustments in the range of 0.9 to 1.1. These transformations were carefully selected to approximate changes in camera perspective while avoiding the introduction of unrealistic artifacts (see Appendix A.3 for full details of the applied transformations). As shown in Table 5.7, these modifications did not yield significant performance improvements compared to the baseline using standard image augmentations, suggesting that the advantages observed with real videos may stem from cues beyond simple geometric or photometric variation. These findings indicate that videos carry rich and distinctive information that plays a crucial role in enhancing adaptation performance—information that cannot be easily replicated through handcrafted augmentations alone.

Chapter 6

Conclusion

In this work, we presented a series of insights regarding the adaptation of visual foundation models to specific target domains. The research led to the development of VESSA (**V**ideo-based **E**fficient **S**elf-**S**upervised **A**daptation for foundation models), a simple, effective, and computationally efficient strategy for unsupervised fine-tuning of vision models using short, object-centric videos. This method is proposed to address the lack of approaches capable of adapting vision foundation models to domain-specific tasks without relying on annotated data.

VESSA operates without the need for labeled data, leveraging the inherent temporal coherence of video sequences. By treating distinct frames of the same video as positive pairs within a contrastive learning framework, the method enables the model to capture richer and more semantically consistent representations. Inspired by successful practices in natural language processing—where continued self-supervised pretraining has become a standard paradigm—we demonstrate that analogous strategies are not only viable in computer vision, but also yield significant performance improvements.

Our empirical results show that VESSA consistently enhances classification accuracy across a variety of domain-specific datasets, while maintaining a lightweight design that can be easily integrated into existing training pipelines. The method proves to be effective across different pretrained vision backbones, underscoring its generality and potential for broad applicability.

To complement our technical contributions, we extend the conclusion by presenting two additional analyses. First, we provide a critical assessment of the current limitations of VESSA, highlighting the factors that may restrict its generalization or performance under specific conditions. Second, we discuss future research directions that naturally emerge from our findings. Rather than merely listing open questions, this section outlines concrete paths for continued investigation, grounded in the empirical patterns and challenges observed during this study.

These analyses are discussed in detail in Sections 6.1 and 6.2, respectively. Together, they provide a broader perspective on the impact, scope, and potential evolution of our approach. Ultimately, we believe that VESSA represents a promising step forward in the pursuit of scalable and annotation-free adaptation strategies for vision foundation

models, fostering their deployment in diverse, real-world scenarios.

6.1 Limitations

This section aims to discuss the primary limitations identified in this work, and consequently, the constraints of the proposed VESSA framework. By acknowledging these limitations, we provide a more realistic perspective on the applicability and scope of our approach, as well as highlight directions for future investigation.

A notable limitation of our approach lies in its susceptibility to forgetting previously acquired knowledge during the fine-tuning process. This phenomenon, often referred to as *catastrophic forgetting*, is a challenge in fine-tuning methods, particularly when models are adapted to new domains without mechanisms to preserve earlier representations. As a result, the benefits of pretraining may be partially compromised, especially when adaptation involves substantial distribution shifts or prolonged training in the target domain.

Another constraint of our methodology arises from the nature of the data used in our experiments. Our approach relies on video sequences that provide multiple viewpoints of the same object, enabling richer visual supervision and more consistent self-supervised learning signals. However, such structured multi-view data is not commonly found in many real-world datasets, potentially limiting the generalizability and scalability of the method.

6.2 Future Work

A promising direction for future research involves the experimental exploration of video behavior in a more comprehensive and systematic manner. By analyzing object-centric and domain-specific video patterns, it may be possible to uncover general principles or recurring visual structures that enhance representation learning. Such investigations could reveal temporal or spatial dynamics in videos that are not captured by current frame selection strategies or adaptation techniques.

The transfer of knowledge from video to image domains remains an underexplored yet highly promising avenue. Object-centric videos, in particular, may serve as a pow-

erful medium for enriching visual representations, especially when adapting to specialized or label-scarce domains. Future studies could focus on optimizing frame selection policies—either through heuristic scoring functions or via trainable frame selection networks—to ensure that the most informative visual content is prioritized during adaptation.

Another promising direction is the development of adaptation strategies that remain effective in less controlled or sparsely-viewed environments. In this context, future work could investigate the use of 3D generative models or simulators to synthesize object-centric videos from various viewpoints. These synthetic video streams may help address the limitations of real-world data availability and provide a controlled environment for training and evaluation.

In addition, further investigation is warranted into alternative loss functions and adaptation objectives for self-supervised fine-tuning. New loss formulations may offer inductive biases that are better suited for domain adaptation, particularly under constraints such as limited training data or significant domain shift. Understanding the impact of these losses on convergence, stability, and downstream performance could lead to more effective and generalizable adaptation pipelines.

Moreover, we suggest exploring the broader use of self-supervised learning (SSL) as a framework for adapting vision foundation models. While SSL has become a cornerstone in representation learning, its potential for domain-specific adaptation remains relatively untapped. Treating SSL as a general-purpose tool for continuous adaptation—rather than solely for pretraining—could open up scalable and annotation-free alternatives to conventional fine-tuning.

Finally, it is essential to investigate the long-term effects of adaptive fine-tuning on vision foundation models, particularly with respect to generalization and catastrophic forgetting. Although domain adaptation can improve performance on target tasks, it may also impair the model's ability to generalize across domains. Future work could examine mechanisms such as regularization, rehearsal-based approaches, or modular adaptation techniques to mitigate forgetting while preserving the foundational versatility of these models.

Collectively, these directions aim not only to extend the capabilities of VESSA, but also to contribute to the broader advancement of scalable, annotation-free adaptation techniques in computer vision. By addressing current limitations and exploring new sources of supervisory signals, future research can unlock more flexible and generalizable strategies for deploying vision foundation models across diverse real-world domains.

References

Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 37–45, 2015.

Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Arthur Aubret, Markus R Ernst, Céline Teulière, and Jochen Triesch. Time to augment self-supervised visual representation learning. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Mohammed Baharoon, Waseem Qureshi, Jiahong Ouyang, Yanwu Xu, Abdulrhman Aljouie, and Wei Peng. Evaluating general purpose vision foundation models for medical image analysis: An experimental study of dinov2 on radiology benchmarks. *arXiv preprint arXiv:2312.02366*, 2023.

Hang Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

Jesimon Barreto, Carlos Caetano, André Araujo, and William R. Schwartz. Vessa: Video-based object-centric self-supervised adaptation for visual foundation models. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025. URL http://arxiv.org/abs/2510.20994.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision (ECCV)*, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Beilei Cui, Mobarakol Islam, Long Bai, and Hongliang Ren. Surgical-dino: adapter learning of foundation models for depth estimation in endoscopic surgery. *International Journal of Computer Assisted Radiology and Surgery*, 19(6):1013–1020, 2024.

Mostafa Dehghani, Alexey Gritsenko, Anurag Arnab, Matthias Minderer, and Yi Tay. Scenic: A jax library for computer vision research and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21393–21398, 2022.

Hao Dong, Moru Liu, Kaiyang Zhou, Eleni Chatzi, Juho Kannala, Cyrill Stachniss, and Olga Fink. Advances in multimodal adaptation and generalization: From traditional approaches to foundation models. *arXiv preprint arXiv:2501.18592*, 2025.

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. In *International Conference on Learning Representations (ICLR)*, 2021.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

Ross Goroshin, Joan Bruna, Jonathan Tompson, David Eigen, and Yann LeCun. Unsupervised learning of spatiotemporally coherent metrics. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4086–4093, 2015.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *CoRR*, abs/2004.10964, 2020. URL https://arxiv.org/abs/2004.10964.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

Rujun Han, Xiang Ren, and Nanyun Peng. DEER: A data efficient language model for event temporal reasoning. *CoRR*, abs/2012.15283, 2020. URL https://arxiv.org/abs/2012.15283.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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.

Xin He, Yushi Chen, Lingbo Huang, Danfeng Hong, and Qian Du. Foundation model-based multimodal remote sensing data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2023.

Jefferson Hernandez, Ruben Villegas, and Vicente Ordonez. Vic-mae: Self-supervised representation learning from images and video with contrastive masked autoencoders. In *European Conference on Computer Vision*, pages 444–463. Springer, 2024.

Edward Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Wei Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2021.

Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 17658–17668, 2024.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer, 2022.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

Samar Khanna, Medhanie Irgau, David B Lobell, and Stefano Ermon. Explora: Parameter-efficient extended pre-training to adapt vision transformers under domain shifts. *arXiv preprint arXiv:2406.10973*, 2024.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Ren Liao, Tony Wei, Zeyi Shen, Sara Beery, Jure Leskovec, Percy Liang, Chelsea Finn Zhang, and Tatsunori Hashimoto. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Tejas D Kulkarni, Ankush Gupta, Catalin Ionescu, Sebastian Borgeaud, Malcolm Reynolds, Andrew Zisserman, and Volodymyr Mnih. Unsupervised learning of object keypoints for perception and control. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Siyuan Li, Luyuan Zhang, Zedong Wang, Di Wu, Lirong Wu, Zicheng Liu, Jun Xia, Cheng Tan, Yang Liu, Baigui Sun, et al. Masked modeling for self-supervised representation learning on vision and beyond. *arXiv preprint arXiv:2401.00897*, 2023.

Zihan Liu, Genta Indra Winata, and Pascale Fung. Continual mixed-language pre-training for extremely low-resource neural machine translation. *CoRR*, abs/2105.03953, 2021. URL https://arxiv.org/abs/2105.03953.

Feng Lu, Xiangyuan Lan, Lijun Zhang, Dongmei Jiang, Yaowei Wang, and Chun Yuan. Cricavpr: Cross-image correlation-aware representation learning for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16772–16782, 2024.

Kangrui Lu, Yuanrun Xu, and Yige Yang. Comparison of the potential between transformer and cnn in image classification. In *ICMLCA 2021; 2nd International Conference on Machine Learning and Computer Application*, pages 1–6. VDE, 2021.

Kevis-Kokitsi Maninis, Kaifeng Chen, Soham Ghosh, Arjun Karpur, Koert Chen, Ye Xia, Bingyi Cao, Daniel Salz, Guangxing Han, Jan Dlabal, Dan Gnanapragasam, Mojtaba Seyedhosseini, Howard Zhou, and André Araujo. TIPS: Text-Image Pretraining with Spatial Awareness. In *ICLR*, 2025.

Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16806–16816, 2023.

Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In *European Conference on Computer Vision (ECCV)*, pages 527–544. Springer, 2016.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*, 2016.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Utku Ozbulak, Hyun Jung Lee, Beril Boga, Esla Timothy Anzaku, Homin Park, Arnout Van Messem, Wesley De Neve, and Joris Vankerschaver. Know your self-supervised learning: a survey on image-based generative and discriminative training. *Transactions on Machine Learning Research*, 2023.

Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2701–2710, 2017.

Jeremy Reizenstein, David Novotny, Deva Ramanan Sun, Federico Tombari, and Andrea Vedaldi. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10901–10911, 2021.

Mohammadreza Salehi, Efstratios Gavves, Cees G. M. Snoek, and Yuki M. Asano. Time does tell: Self-supervised time-tuning of dense image representations. *ICCV*, 2023.

Linus Scheibenreif, Michael Mommert, and Damian Borth. Parameter efficient self-supervised geospatial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27841–27851, 2024.

Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, 2015.

Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12345–12354. IEEE, 2023. doi: 10.1109/CVPR.2023.01234.

Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403, 2021.

Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, 2016.

Yang Zhang, Caiqi Liu, Mujiexin Liu, Tianyuan Liu, Hao Lin, Cheng-Bing Huang, and Lin Ning. Attention is all you need: utilizing attention in ai-enabled drug discovery. *Briefings in bioinformatics*, 25(1), 2023.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *International Journal of Machine Learning and Cybernetics*, pages 1–65, 2024.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*, 2022.

Xuechao Zou, Shun Zhang, Kai Li, Shiying Wang, Junliang Xing, Lei Jin, Congyan Lang, and Pin Tao. Adapting vision foundation models for robust cloud segmentation in remote sensing images. *arXiv preprint arXiv:2411.13127*, 2024.

Appendix A

Additional Information

A.1 MVImagnet Classes

The table below lists all object categories included in the MVImageNet dataset, totaling 238 labeled classes. These categories encompass a wide range of common and specialized items, including household objects, tools, animals, food items, and toys.

1. bag
2. bottle
3. wahser
4. vessel
5. train
6. telephone
7. table
8. stove
9. sofa
10. skateboard
11. rifle
12. pistol
13. remote control
14. printer
15. flowerpot
16. pillow
17. piano
18. mug
19. motorcycle
20. microwave
21. microphone
22. mailbox
23. loudspeaker
24. laptop
25. lamp
26. knife
27. pot
28. helmet
29. guitar
30. bookshelf
31. faucet
32. earphone
33. display
34. dishwasher
35. computer keyboard
36. clock
37. chair
38. car
39. cap
40. can
41. camera
42. cabinet
43. bus
44. bowl
45. bicycle

46. bench
47. bed
48. bathtub
49. basket
50. ashcan
51. airplane
52. umbrella
53. plush toy
54. toy figure
55. towel
56. toothbrush
57. toy bear
58. toy cat
59. toy bird
60. toy insect
61. toy cow
62. toy dog
63. toy monkey
64. toy elephant
65. toy fish
66. toy horse
67. toy sheep
68. toy mouse
69. toy tiger
70. toy rabbit
71. toy dragon
72. toy snake
73. toy chook
74. toy pig
75. rice cooker
76. pressure cooker
77. toaster
78. dryer
79. battery
80. curtain
81. blackboard eraser
82. bucket
83. calculator
84. candle
85. cassette
86. cup sleeve
87. computer mouse
88. easel
89. fan
90. cookie
91. fries
92. donut
93. coat rack
94. guitar stand
95. can opener
96. flashlight
97. hammer
98. scissors
99. screw driver
100. spanner
101. hanger
102. jug
103. fork
104. chopsticks
105. spoon
106. ladder
107. ceiling lamp
108. wall lamp
109. lamp post
110. light switch
111. mirror
112. paper box
113. wheelchair
114. walking stick
115. picture frame
116. shower
117. toilet
118. sink
119. power socket
120. Bagged snacks
121. Tripod
122. Selfie stick
123. Hair dryer

124. Lipstick
125. Glasses
126. Sanitary napkin
127. Toilet paper
128. Rockery
129. Chinese hot dishes
130. Root carving
131. Flower
132. Book
133. Pipe PVC Metal pipe
134. Projector
135. Cabinet Air Conditioner
136. Desk Air Conditioner
137. Refrigerator
138. Percussion
139. Strings
140. Wind instruments
141. Balloons
142. Scarf
143. Shoe
144. Skirt
145. Pants
146. Clothing
147. Box
148. Soccer
149. Roast Duck
150. Pizza
151. Ginger
152. Cauliflower
153. Broccoli
154. Cabbage
155. Eggplant
156. Pumpkin
157. winter melon
158. Tomato
159. Corn
160. Sunflower
161. Potato
162. Sweet potato
163. Chinese cabbage
164. Onion
165. Momordica charantia
166. Chili
167. Cucumber
168. Grapefruit
169. Jackfruit
170. Star fruit
171. Avocado
172. Shakyamuni
173. Coconut
174. Pineapple
175. Kiwi
176. Pomegranate
177. Pawpaw
178. Watermelon
179. Apple
180. Banana
181. Pear
182. Cantaloupe
183. Durian
184. Persimmon
185. Grape
186. Peach
187. power strip
188. Racket
189. Toy butterfly
190. Toy duck
191. Toy turtle
192. Bath sponge
193. Glove
194. Badminton
195. Lantern
196. Chestnut
197. Accessory
198. Shovel
199. Cigarette
200. Stapler

201. Lighter
202. Bread
203. Key
204. Toothpaste
205. Swin ring
206. Watch
207. Telescope
208. Eggs
209. Bun
210. Guava
211. Okra
212. Tangerine
213. Lotus root
214. Taro
215. Lemon
216. Garlic
217. Mango
218. Sausage
219. Besom
220. Lock
221. Ashtray
222. Conch
223. Seafood
224. Hairbrush
225. Ice cream
226. Razor
227. Adhesive hook
228. Hand Warmer
229. Thermometer
230. Bell
231. Sugarcane
232. Adapter(Water pipe)
233. Calendar
234. Insecticide
235. Electric saw
236. Inflator
237. Ironmongery
238. Bulb

A.2 CO3D Classes

The CO3D dataset consists of 50 common object categories captured in real-world scenarios. These categories cover a broad spectrum of everyday items and are listed below.

1. backpack
2. baseball bat
3. baseball glove
4. basket
5. bench
6. bicycle
7. book
8. bottle
9. bowl
10. cake
11. car
12. carrot
13. cell phone
14. chair
15. clock
16. couch
17. cup
18. dining table
19. dog
20. donut
21. elephant
22. fire hydrant
23. frisbee
24. giraffe

25. hair drier
26. handbag
27. hot dog
28. hydrant
29. keyboard
30. kite
31. laptop
32. microwave
33. motorcycle
34. mouse
35. orange
36. oven
37. parking meter
38. pizza
39. plant
40. refrigerator
41. remote
42. sandwich
43. scissors
44. sink
45. skateboard
46. spoon
47. sports ball
48. suitcase
49. teddy bear
50. toaster
51. toilet

A.3 Augmentation Pipeline Details

To support effective self-supervised training, our method relies on a carefully designed data augmentation pipeline. This pipeline follows the general principles of multi-crop augmentation used in DINO [Caron et al., 2021], generating two global views and multiple local crops per input image. Each view is subjected to a specific combination of transformations intended to induce invariance to appearance-level variations while preserving semantic content. The full set of transformations applied is described below.

Global crops: Two global crops are sampled from the image with scale ranges between $(0.4, 1.0)$ and resized to 224×224 pixels. Each crop is augmented independently with a distinct sequence of transformations:

- **Transformation view 1:**
  - Random horizontal flip (probability 0.5),
  - Color jitter (strength 0.8),
  - Random grayscale conversion (probability 0.2),
  - Gaussian blur (probability 1.0).
- **Transformation view 2:**
  - Random horizontal flip (probability 0.5),
  - Color jitter (strength 0.8),

  - Random grayscale conversion (probability 0.2),
  - Gaussian blur (probability 0.1),
  - Solarization (probability 0.2).

Local crops: For each image, a set of u local crops is sampled with scale range $(0.05, 0.25)$ and resized to 96×96 pixels. Each local crop is transformed independently using the following sequence:

- Color jitter with parameters: brightness 0.4, contrast 0.4, saturation 0.2, hue 0.1 (strength 0.8),
- Random grayscale conversion (probability 0.2),
- Gaussian blur (probability 0.5).

Motion-inspired augmentations: To simulate viewpoint variation without relying on video frames, we applied the following geometric and photometric transformations to one of the global crops:

- Translation of up to 10% of image width and height,
- Rotation up to 10°,
- Scaling variation up to 5%,
- Brightness shift of 0.1,
- Contrast adjustment in the range $[0.9, 1.1]$.

These transformations were carefully selected to approximate camera motion while minimizing unrealistic artifacts. Despite their similarity to natural viewpoint changes, our experiments show that they do not fully replicate the benefits of using real video-based inputs.