

A Technical Appendices and Supplementary Material

In this Technical Appendix and Supplementary Material, we present the complementary experiments of the paper. These evaluations include the main experiments with confidence intervals for the primary results, demonstrating the statistical significance of the observed differences. We further analyze the input differences between DINO and VESSA, and investigate the impact of applying stronger image augmentations, aiming to reduce the gap between static images and the visual variability observed in video data. We also present an additional study examining the extent of catastrophic forgetting when adapting visual foundation models with VESSA, highlighting its impact on general-purpose performance. Finally, we provide a detailed analysis of the training cost associated with VESSA, offering quantitative insights into its computational efficiency and practical feasibility.

The main results with confidence intervals are shown in Tables 1 and 2, which report the top-performing models along with confidence intervals to highlight the significance of the performance differences. The results suggest that VESSA achieves significantly better performance in several cases. We employed an unpaired Student’s *t*-test with a 90% confidence level. We report confidence intervals for the main comparisons to highlight cases where the differences were statistically significant. For the CO3D dataset, the variation in accuracy across runs was 0.52 for DINO, 0.56 for DINOv2, and 1.03 for TIPS. In all cases, the differences were statistically significant when compared to the second-best performing method, which in this case was ExPLoRA—a method also based on video data. In contrast, on the MVImageNet dataset, the variations were 1.11 for DINO, 1.08 for DINOv2, and 1.71 for TIPS. In this scenario, non-overlapping confidence intervals between the DINO-based baseline and VESSA (ours) indicate a statistically significant difference. In the case of DINOv2 pretrained on MVImageNet, a noticeable overlap between the two video-based methods can be observed.

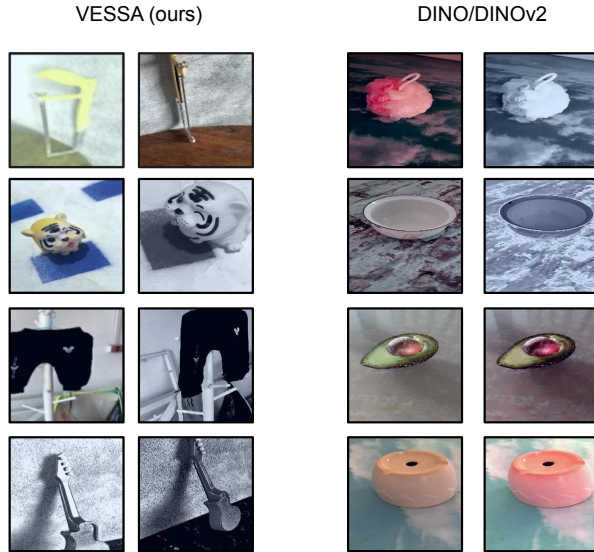


Figure 1: Example frames from the MVImageNet dataset illustrating the differences between global crop input pairs used for the teacher and student networks during training with DINO and VESSA (ours). Our method, VESSA, introduces substantially greater variability in the appearance of the evaluated object. 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.

Approximate video-like image transformations. 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

Table 1: **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 \pm 0.84	89.64 \pm 0.47	—
VESSA (ours)	85.03 \pm 0.52	91.85 \pm 0.56	70.56 \pm 1.03

Table 2: **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 \pm 1.03	96.15 \pm 0.87	—
VESSA (ours)	92.51 \pm 1.11	96.01 \pm 1.08	80.54 \pm 1.71

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. As shown in Table 3, 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.

The data augmentation pipeline consists of two global views and multiple local crops, each subjected to a specific set of transformations, as detailed below.

- **Global crops:** Two crops are sampled with scale ranges between (0.4, 1.0). The transformations applied to these global crops are as follows:
 - **Transformation view 1:** horizontal flip (probability 0.5), color jitter (strength 0.8), grayscale conversion (probability 0.2), and Gaussian blur (probability 1.0).
 - **Transformation view 2:** horizontal flip (probability 0.5), color jitter (strength 0.8), grayscale conversion (probability 0.2), Gaussian blur (probability 0.1), and solarization (probability 0.2).
- **Local crops:** A set of u local crops per image are sampled with a scale range (0.05, 0.25) defined by the configuration and resized to 96×96 pixels. Each local crop undergoes the following transformations independently:
 - Color jitter with parameters (strength 0.8, brightness 0.4, contrast 0.4, saturation 0.2, hue 0.1).
 - Grayscale conversion (probability 0.2).
 - Gaussian blur (probability 0.5).

Impact of using video-based inputs. To highlight the differences between the view generation strategies employed by VESSA and those used in DINO, we present illustrative examples of view pairs from both approaches. It is important to note that in both VESSA and DINO, the same groups of transformations are applied independently to each view. To assess the impact of video-based inputs on representation learning, we analyze input variability by contrasting the frame selection strategy used in VESSA with the standard augmentation-based sampling in DINO and DINOv2. As shown in Figure 1, 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

Table 3: 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

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.

Impact of catastrophic forgetting and cross-dataset generalization. We also investigate the impact of domain-specific adaptation with VESSA on the original pretraining task. To this end, we evaluate the performance of DINO, DINOv2, and TIPS on ImageNet classification using a k-nearest neighbors (KNN) classifier, both in their pretrained form and after adaptation on CO3D and MVImageNet. As shown in Table 4, while the pretrained models achieve competitive accuracy on ImageNet, their performance drops drastically once adapted with VESSA on either domain. This result confirms the presence of catastrophic forgetting and underscores the infeasibility of applying the adapted models in a general-purpose setting. Nonetheless, this outcome aligns with the intended design of VESSA: the method is tailored to specialize visual foundation models for unsupervised adaptation in a target domain, where it delivers strong performance despite losing generality.

Table 4: Effect of catastrophic forgetting on ImageNet classification after VESSA adaptation.

Model	DINO	DINOv2	TIPS
Pretrained	76.10	82.10	80.00
VESSA (CO3D)	15.46	17.15	18.10
VESSA (MVImagnet)	15.68	16.78	17.10

Moreover, we conducted an experiment in which training was 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, 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 the presence of catastrophic forgetting and limited generalization capabilities when the model is exposed to a distribution shift, even when trained on a diverse and temporally rich video corpus.

Table 5: 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	DINOv2
Pretrained	78.86	87.86
Cross dataset	74.40	80.36

Training cost and computational efficiency. We also analyze the computational requirements of adapting visual foundation models with VESSA to assess its practical feasibility. For the Co3D dataset (20, 412 training pairs and 4, 535 test samples) using a ViT-Base backbone, the adaptation stage required only 1.97 hours of training, corresponding to 7.04 core-hours on a TPU v3-8, which consists of 4 TPUs, each with 2 cores (total of 8 cores). The process ran for 20 epochs, processing 407, 040 examples at a throughput of 135 images per second (16.88 images per second per core), with a total energy consumption of 13.86 kWh and an estimated carbon footprint of 5.55kg CO₂. Importantly, VESSA adapts a pre-trained model rather than training from scratch, and the minor additional overhead introduced by processing paired video frames (or local crops) is negligible due to

offline video decoding. This efficiency contrasts sharply with the full pretraining of models such as DINO or DINOv2, which require thousands of GPU-hours and result in CO₂ emissions on the order of tons.