MULTIVIEW EQUIVARIANCE IMPROVES 3D UNDERSTANDING WITH MINIMAL FEATURE FINETUNING

Anonymous authors
Paper under double-blind review

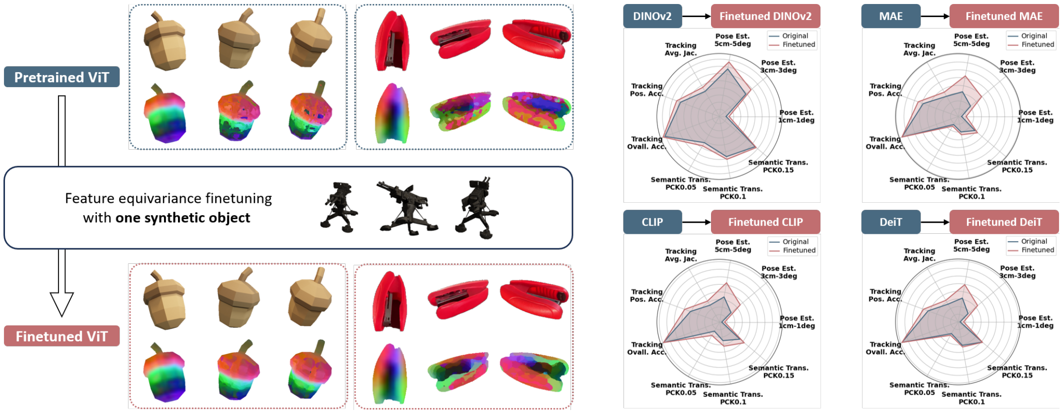

Figure 1: **Improving 3D understanding through finetuning on feature equivariance. Left:** finetuning feature equivariance on one synthetic object can already enhance the vision transformer's ability to generate better 3D feature correspondences on general objects. **Right:** This improvement further leads to superior performance across multiple 3D tasks, including pose estimation, video tracking, and semantic correspondence.

ABSTRACT

Vision foundation models, particularly the ViT family, have revolutionized image understanding by providing rich semantic features. However, despite their success in 2D comprehension, their abilities on grasping 3D spatial relationships are still unclear. In this work, we evaluate and enhance the 3D awareness of ViT-based models. We begin by systematically assessing their ability to learn 3D equivariant features, specifically examining the consistency of semantic embeddings across different viewpoints. Our findings indicate that improved 3D equivariance leads to better performance on various downstream tasks, including pose estimation, tracking, and semantic transfer. Building on this insight, we propose a simple yet effective finetuning strategy based on 3D correspondences, which significantly enhances the 3D understanding of existing vision models. Remarkably, even finetuning on a single object for just one iteration results in substantial performance gains. All code and resources will be made publicly available to support further advancements in 3D-aware vision models.

1 INTRODUCTION

Common camera imaging systems struggle to depict the 3D world due to the limitation of capturing only a single perspective at any given moment. In contrast, human perceptual capabilities exhibit a remarkable trait known as view equivariance Köhler (1967); Koffka (2013); Wilson & Farah (2003), allowing us to robustly understand 3D spatial relationships, as seen in tasks ranging from basic object recognition Vetter et al. (1995); DiCarlo & Cox (2007) to more complex processes like mental rotation and simulation Stewart et al. (2022).

Current large vision models, however, are primarily trained on 2D images, owing to the ease of data acquisition and annotation in 2D. Consequently, their performance is typically evaluated on 2D tasks Amir et al. (2021); Hedlin et al. (2023); Tang et al. (2023); Zhang et al. (2023). This raises critical questions: *To what extent do these models possess an inherent awareness of 3D structures? How does this awareness impact their performance on image-based 3D vision tasks? And, can we further enhance the 3D awareness of these vision foundation models?*

Many image-based 3D scene understanding and content generation tasks depend heavily on large 2D vision models, underscoring the importance of investigating these questions. Existing works have begun to explore this area in task-specific contexts. For example, DietNeRF Jain et al. (2021) finds that CLIP Radford et al. (2021) demonstrates higher feature similarities between views from the same scene than from different scenes, which aids 3D reconstruction. LeRF Kerr et al. (2023) shows that regularizing CLIP with DINO Caron et al. (2021) features improves 3D feature distillation from multiple views. However, these studies are tied to specific tasks such as feature distillation. El Banani et al. (2024) probes the multi-view consistency of ViTs by evaluating them on the NAVI and ScanNet datasets. However, the limited size of these datasets makes it challenging to draw comprehensive conclusions.

To address the first question, *how well do vision models understand 3D structures*, we present a comprehensive study of the 3D awareness of large 2D vision models. Specifically, we investigate the *view equivariance* of latent features—i.e., the consistency of multi-view 2D image features representing the same 3D point across different views. Using off-the-shelf multiview correspondences rendered from Objaverse Deitke et al. (2023) (synthetic) and MVImgNet Yu et al. (2023) (real-world), we find that current large vision models do exhibit some degree of view-consistent feature generation, with DINOv2 demonstrating the strongest performance.

To answer the second question, *how does this awareness influence performance in image-based 3D vision tasks*, we find that the quality of 3D equivariance is strongly correlated with performance on three downstream tasks requiring 3D understanding: pose estimation, video tracking, and semantic correspondence. Consistent with previous findings Örnek et al. (2023); Tumanyan et al. (2024); Zhang et al. (2023), DINOv2 Oquab et al. (2023) excels in these tasks, showcasing its robust capability in 3D vision.

Finally, to address the third question, *can we improve the 3D awareness of vision foundation models*, we propose a simple yet effective method to enhance the view equivariance of 2D foundation models, thereby significantly improving their 3D understanding. During training, we randomly select two different views of the same object from Objaverse and sample corresponding pixels. We apply the SmoothAP Brown et al. (2020) loss to enforce feature similarity between these corresponding pixels. This finetuning process, requiring only 10K iterations with LoRA and an additional convolutional layer of a Vision Transformer (ViT), significantly improves the performance of all tested models on 3D tasks. For instance, DINOv2 gains improvements of **9.58** (3cm-3deg in pose estimation), **5.0** (Average Jaccard in tracking), and **5.06** (PCK@0.05 in semantic correspondence). Surprisingly, even finetuning on a single multi-view pair sampled from one object for just one iteration yields notable gains in 3D understanding. In such cases, DINOv2's performance improves by **4.85**, **3.55**, and **3.47** for 3cm-3deg (pose estimation), Average Jaccard (tracking), and PCK@0.05 (semantic correspondence), respectively.

To summarize, our key contributions are: (i) We conduct a comprehensive evaluation of 3D equivariance capabilities in 2D vision foundation models. (ii) We demonstrate that the quality of 3D equivariance is closely tied to performance on three downstream tasks that require 3D understanding: pose estimation, video tracking, and semantic correspondence. (iii) We propose a simple but effective finetuning method that substantially improves the 3D understanding of 2D foundation models, leading to marked performance gains across all evaluated tasks.

2 EVALUATION OF MULTIVIEW FEATURE EQUIVARIANCE

To assess how effectively current vision transformers capture 3D understanding, we introduce a 3D equivariance evaluation benchmark focused on the quality of correspondences between 2D points across different views for the same object. Additionally, we present three well-established application tasks that rely on 3D correspondence, demonstrating a strong correlation between the quality of 3D equivariance and downstream task performance. We evaluate five state-of-the-art vision transformers: DINOv2 Oquab et al. (2023), DINOv2-Reg Darcet et al. (2023), MAE He et al. (2022a), CLIP Radford et al. (2021) and DeiT Touvron et al. (2022), extracting their final-layer features with

L2 normalization. For DINOv2, we use the *base* model; results for other variants are provided in the supplementary material.

To evaluate 3D equivariance, we utilize rendered or annotated multiview correspondences from **Objaverse** Deitke et al. (2023) and **MVImgNet** Yu et al. (2023), covering both synthetic and real images. For Objaverse, we randomly select 1,000 objects from the Objaverse repository, rendered across 42 uniformly distributed camera views, producing 42,000 images. Dense correspondences are computed for each object across every unique ordered pair of views, resulting in 1.8 billion correspondence pairs for evaluation. Similarly, 1,000 objects are randomly drawn from MVImgNet, yielding 33.3 million annotated correspondence pairs for evaluation. Since MVImgNet employs COLMAP to reconstruct 3D points, it provides sparser correspondences compared to Objaverse.

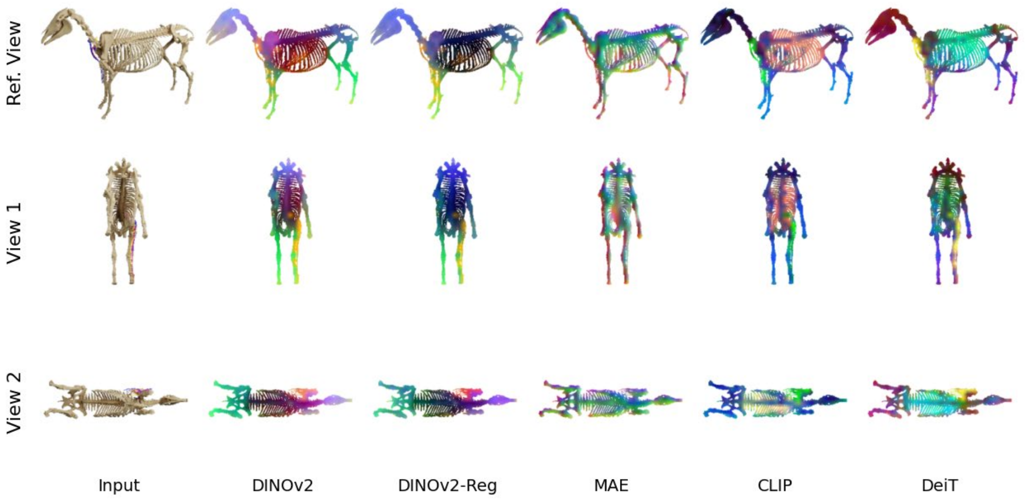

Figure 2: **Feature visualizations of different models.** The sample image is rendered from Objeverse. Colors are computed from the high-dimensional features using PCA. We can see that MAE struggles to distinguish different parts of the content (*e.g.*similar features between head and body). Both CLIP and DeiT produce inconsistent features for the chest region between View 1 and View 2. DINOv2 gives the best correspondence.

Metric and Results We propose the **Average Pixel Error%** (APE), a metric that quantifies the average distance between predicted and ground-truth pixel correspondences, normalized by the length of the shortest image edge. The predicted correspondence is determined by identifying the nearest neighbor in the second view, given a reference point feature in the first view. APE for Objaverse is shown in Figure 3, where APE is plotted on the x-axis, meaning lower values (towards the left) indicate better performance. APE and PCDP for MVImgNet are plotted on Figure 5's y-axis with hollow circle ○ and striped bar ▨ representing the evaluted pretrained models (fine-tuning results will be discussed later). **Percentage of Correct Dense Points%** (PCDP) is a metric designed to evaluate dense correspondences, similar to Percentage of Correct Keypoints% (PCK). It is reported at various thresholds (5%, 10%, and 20% of the shortest image edge). We can see that DINOv2 and its registered version outperform other vision transformers, highlighting DINOv2's superior capability for 3D equivariance. In Figure 15, we provide feature visualizations using PCA, where DINOv2 again demonstrates the best multiview feature consistency.

2.1 FEATURE EQUIVARIANCE CORRELATES TO CERTAIN TASK PERFORMANCES

3D Equivariance itself is not interesting unless it can be used. Below, we will talk about three mature downstream applications that require 3D equivariance capability, and show a correlation between the quality of 3D equivariance and the downstream applications.

2.1.1 TASK DEFINITIONS

One-Shot Object Pose Estimation In the one-shot pose estimation task, we assume the availability of either a video sequence or a 3D mesh of the target object and aim to estimate its pose

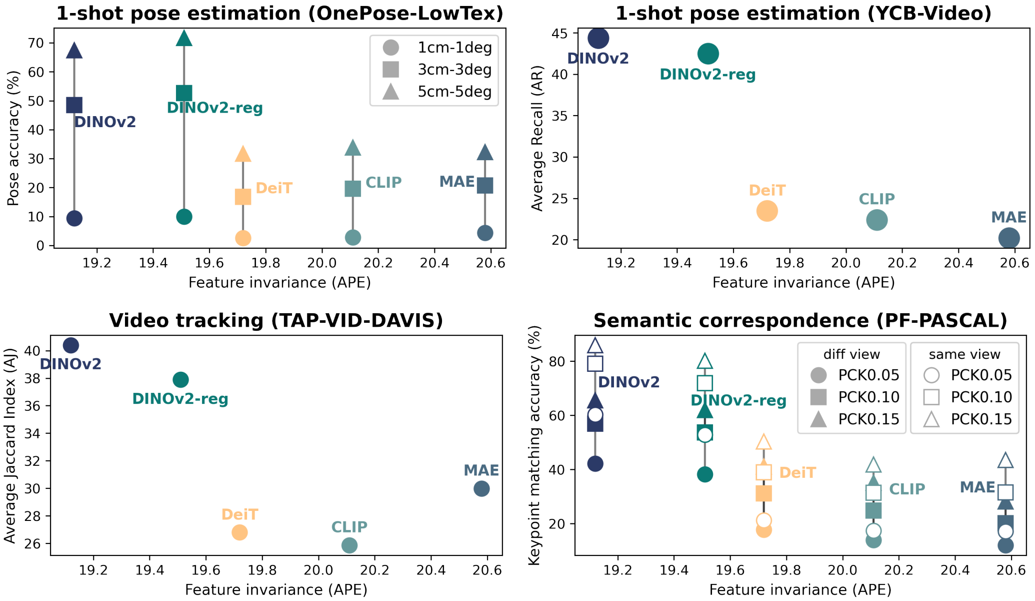

Figure 3: **Correlation between multiview feature equivariance and the task performances.** Along the horizontal axis, lower APE indicates better feature equivariance, while the vertical axis reflects higher task performance across all four plots. The data points align roughly along the diagonal from the top left to the bottom right, suggesting a strong correlation between improved feature equivariance and better task performance.

in arbitrary environments. During training, we store the extracted dense 2D image features from all rendered or annotated views as a database. During inference, correspondences between the input image and the stored features are computed, allowing us to match 2D keypoints from the input image to their 3D counterparts in the database. To robustly estimate the object's pose from these 2D-3D correspondences, we employ RANSAC Fischler & Bolles (1981) PnP (Perspective-n-Point). Specifically, we uniformly sample points across the image using stratified sampling with a stride of 4. The image is resized to 512×512, and the number of iterations in RANSAC PnP is set to 10,000, with a threshold of 8.

We evaluate on the OnePose-LowTexture and YCB-Video datasets. OnePose-LowTexture, proposed in OnePose++ He et al. (2022b), assesses one-shot pose estimation for textureless objects and includes 40 low-textured household items. Each object is captured in two videos with different backgrounds: one for reference and one for testing, forming a one-shot estimation scenario. Following OnePose++ He et al. (2022b), we evaluate pose accuracy using thresholds ranging from 1cm-1deg, 3cm-3deg to 5cm-5deg. The YCB-Video dataset Xiang et al. (2017) consists of 21 objects and 92 RGB-D video sequences with pose annotations, along with CAD models of the evaluated objects for one-shot generalization. We construct a database by rendering the target object from 96 icospherical viewpoints. We follow the current literature to report Average Recall (AR) for Visible Surface Discrepancy (VSD), Maximum Symmetry-Aware Surface Distance (MSSD), and Maximum Symmetry-Aware Projection Distance (MSPD), with further details available in Hodaň et al. (2020).

Video Tracking For video tracking, given the reference frame, we identify corresponding points in other frames by computing cosine similarities between the dense features of the target object. To improve robustness and accuracy, we follow the process in DINO-Tracker Tumanyan et al. (2024), which applies a softmax operation within the neighborhood of the location with highest similarity.

We evaluate the models on the TAP-Vid-DAVIS Doersch et al. (2022) dataset, a benchmark designed for testing video tracking in complex, real-world scenarios. The dataset includes sequences with challenging object motions, varying appearances, occlusions, and dynamic backgrounds, providing an ideal setting for assessing temporal consistency and robustness. Performance is measured using commonly applied metrics Tumanyan et al. (2024), including the Average Jaccard Index (AJ), Position Accuracy (δ^x_{avg}), and Occlusion Accuracy (OA).

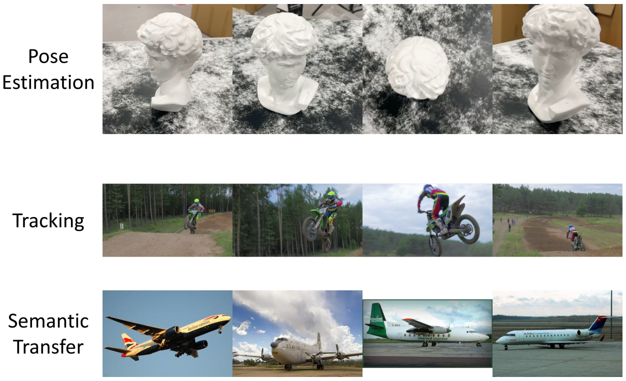

Figure 4: **Illustration of different types of correspondence tasks evaluated in our work.** (Top) Pose estimation requires matching points between different views of the same rigid object. (Middle) Video tracking involves tracking points on non-rigid or articulated objects across frames. (Bottom) Semantic transfer aims to find corresponding points between different instances of similar semantic categories, such as different aircraft models.

Semantic Correspondence In the semantic correspondence task, we utilize feature correspondences to establish precise keypoint matches between images captured from different instances from the same category. Following the method in Zhang et al. (2023), for a given reference keypoint, we identify the best match by selecting the location that exhibits the highest cosine similarity between the dense feature representations

We use the PF-PASCAL Ham et al. (2017) dataset as our evaluation benchmark. This dataset typically consists of image pairs taken from the same viewpoint, but we additionally report the result by shuffling the image pairs to include different viewpoints, thereby increasing the challenge. This modification introduces greater difficulty in matching keypoints, as it requires the model to find correspondences between views with significant perspective variation, emphasizing the model's 3D spatial understanding capabilities. We follow standard practice to report results using PCK0.05, PCK0.10, and PCK0.15 as evaluation metrics.

2.1.2 ON THE CHOICE OF THREE TASKS

Correspondence estimation is a fundamental component of 3D vision understanding, underlying key tasks such as epipolar geometry, stereo vision for 3D reconstruction, and optical flow or tracking to describe the motion of a perceived 3D world. Stereo cameras, and even human perception, rely on disparity maps—effectively, correspondences between projected 3D parts to understand depth and spatial relationships. The three tasks we evaluated—pose estimation, video tracking, and semantic correspondence—were intentionally selected to cover diverse aspects of correspondence estimation, ranging from simpler to more complex scenarios:

- Pose Estimation examines correspondences within the same instance under rigid transformations ($SE(3)$);
- Video Tracking extends this to correspondences for the same instance under potential non-rigid or articulated transformations, such as humans or animals in motion;
- Semantic Correspondence requires correspondences across different instances with similar semantics, often under arbitrary transformations.

An qualitative illustration of these correspondence types is shown in Figure 4.

2.1.3 RESULTS AND FINDINGS

Quantitative results are presented in Figure 3, where the y-axis in each graph shows the performance of the vision models. DINOv2 consistently outperforms all other models across all three tasks, in alignment with the rankings for 3D equivariance on the x-axis. There is a clear correlation between the quality of 3D equivariance and performance on the downstream tasks: methods with lower APE tend to perform better across all tasks, clustering towards the top-left of the graphs.

3 FEATURE FINETUNING WITH MULTIVIEW EQUIVARIANCE

Given the correlation between the multiview equivariance of network features and task performances, we naturally come up with a question: *Can we finetune the networks on feature equivariance to improve their 3D understanding and achieve better task performances?*

Finetuning method The high-level intuition of improving the multiview equivariance of the network features is to enforce the similarity between features of corresponding pixels in 3D space. Toward this goal, we experiment with multiple strategies including different training objectives and network architectures.

For the training loss, rather than employing a conventional contrastive loss, we opted for the SmoothAP Brown et al. (2020) loss, which demonstrated superior performance. While contrastive loss can help align the features of corresponding pixels, it relies on a predefined fixed margin for positive and negative samples, which is ad hoc and often suboptimal. In contrast, SmoothAP optimizes a ranking loss directly, leading to an improved average precision for feature retrieval between corresponding pixels. We also experimented with the differentiable nearest neighbor and Procrustes alignment loss from Li et al. (2022), but these did not outperform SmoothAP. Detailed ablation results are given in Section 4.3.

In terms of architecture, besides the common practice of using LoRA to finetune large foundation models, we introduced a single convolutional layer with a kernel size of 3 and a stride of 1. The motivation behind this addition is rooted in the observation that ViT-family models process image tokens as patches, resulting in much lower-resolution feature maps (e.g., 14x smaller in DINOv2). The standard approach to obtain high-resolution per-pixel features is to apply linear interpolation. Consequently, it is beneficial to explicitly exchange information between neighboring patches before interpolation to achieve more accurate results. More details and ablation results are given in Section 4.1.

During training, at each iteration, we randomly select two different views of the same object from Objaverse and sample a set of corresponding pixels. The object is also sampled from a 10K random subset of Objavserse. We train the model for 10K iterations with the AdamW optimizer of learning rate 1e-5 and weight decay 1e-4. In the supplementary, we also show that our finetuning method is insensitive to the learning rate.

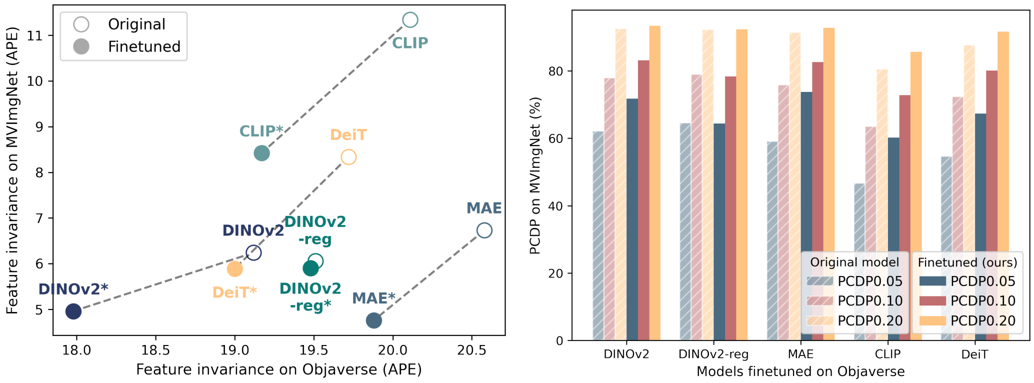

Figure 5: **Generalization from synthetic images (Objaverse) to real images (MVImgNet). Left:** Data points roughly around the diagonal from the bottom left to the upper right indicate the correlation between the APE tested on the two datasets. The * next to the model name means it is finetuned. All finetuning is done on Objaverse with only synthetic data. **Right:** Finetuned on Objaverse, the feature equivariance of the model (measured in PCDP) improves on MVImgNet.

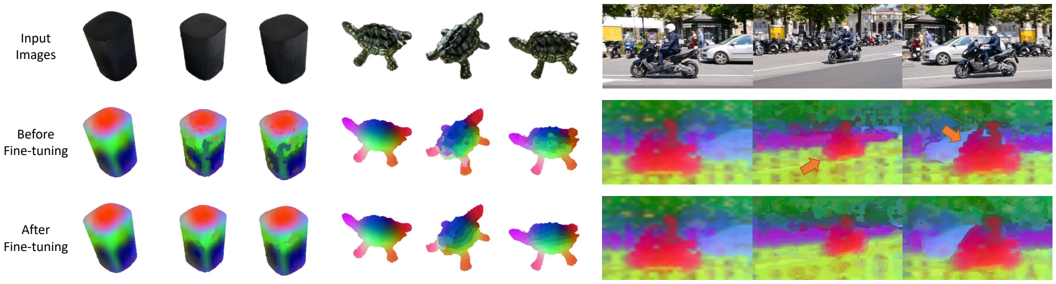

Figure 6: **Feature visualization of DINOv2 before and after finetuning on MVImgNet objects (left two) and TAP-VID-DAVIS scenes (right one).** For each example, we select three different views. The first column provides a reference color produced by PCA, while the second and third columns show the predicted feature correspondences. Our finetuned model demonstrates reduced noise and smoother feature boundaries, particularly noticeable in the reduction of jagged edges, indicating improved feature consistency across views.

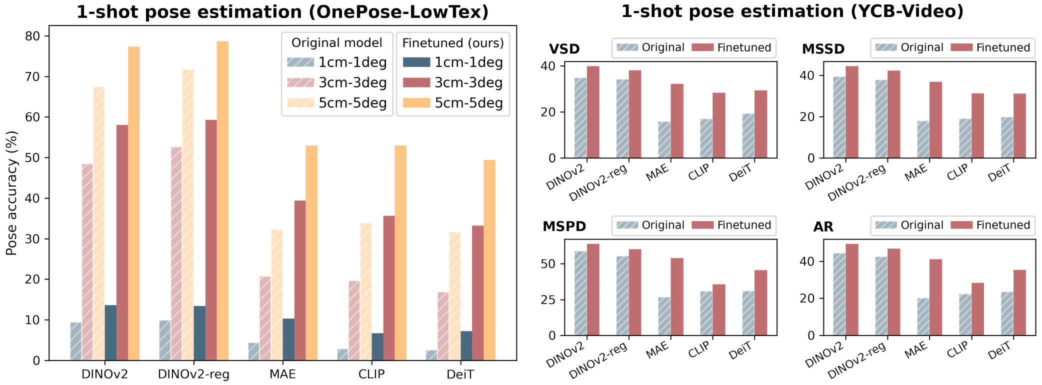

Figure 7: **One-shot pose estimation results before and after feature equivariance finetuning.** The feature finetuning consistently improves the performance of all ViTs.

3.1 IMPROVED FEATURE EQUIVARIANCE WITH GENERALIZATION

Figure 5 illustrates the performance of various models before and after finetuning. After finetuning on Objaverse, all models show improved 3D equivariance on both Objaverse (synthetic) and MVImgNet (real-world). This demonstrates the capacity of vision foundation models to perform sim-to-real transfer, as finetuning on synthetic Objaverse objects results in enhanced performance on the real-world MVImgNet dataset. Additionally, the performance on the two datasets is correlated, with data points roughly aligning along the diagonal, indicating that improvements in synthetic environments translate well to real-world settings. DINOv2 stands out as the best model. We also compare the feature visualizations before and after finetuning in Figure 6, from which we can see that after finetuning the model produces more consistent features with less noise.

3.2 IMPROVED TASK PERFORMANCES

One-shot Object Pose Estimation Figure 7 shows the performance of pose estimation on the OnePose-LowTex and YCB-Video datasets before and after fine-tuning. As illustrated, all Vision Transformers (ViTs) exhibit noticeable improvements after being fine-tuned on synthetic Objaverse data. For instance, the best-performing model, DINOv2-Reg, improves by **3.46**, **6.67**, and **6.92** for the 1cm-1deg, 3cm-3deg, and 5cm-5deg thresholds, respectively. Additionally, models that performed weaker before fine-tuning show larger gains. For example, DeiT improves by **4.65**, **16.39**, and **17.76**. Similar trends are observed for the YCB-Video dataset, where models like MAE, initially the weakest, show substantial improvement after fine-tuning.

Video Tracking Similarly, in the video tracking task, we observe consistent improvements across all ViTs after fine-tuning, as shown in Figure 8. The top-performing model, DINOv2, achieves improvements of **6.45**, **5.73**, and **2.69** in Average Jaccard (AJ), Position Accuracy (δ^x_{avg}), and Occlusion Accuracy (OA), respectively.

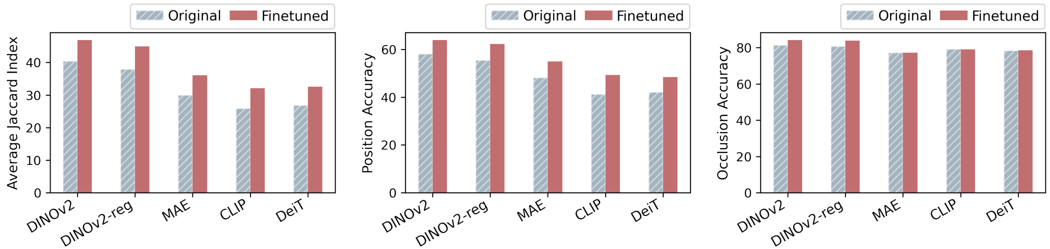

Figure 8: **Video tracking results before and after feature equivariance finetuning.** The feature finetuning consistently improves the performance of all ViTs.

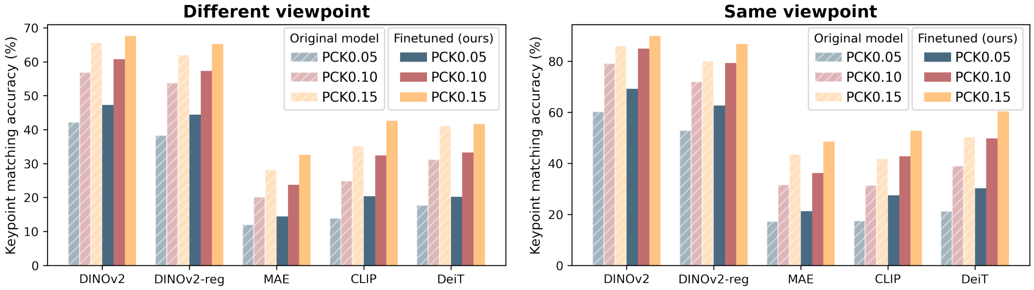

Figure 9: **Semantic correspondence results before and after feature equivariance finetuneing.** The feature finetuning consistently improves the performance of all ViTs.

Semantic Correspondence In the semantic correspondence task, shown in Figure 9, DINOv2 exhibits improvements of **5.06**, **3.86**, and **1.98** for PCK@0.05, PCK@0.10, and PCK@0.15, respectively. Notably, we find that fine-tuned models show enhanced understanding of keypoint semantics across different instances, even from the same viewpoint. This insight suggests that 3D equivariance contributes to a better understanding of fine-grained semantics, despite not being explicitly trained or finetuned for that purpose.

We also compared with FiT Yue et al. (2024) and DUSt3R Wang et al. (2024), while their performance are much worse than ours. Detailed quantitative results including FiT and DUSt3R on all these tasks are available in the supplementary materials.

3.3 EXTREMELY FEW-SHOT FINETUNING

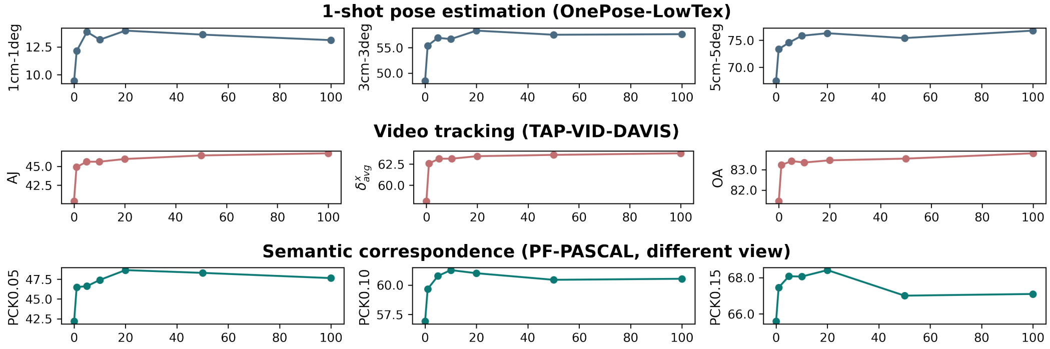

Figure 10: **Finetuned performances *w.r.t.* #training objects.** We evaluate the performances of the DINOv2 model finetuned with 0, 1, 5, 10, 20, 50, 100 objects on the three tasks.

Training with Only One Object We plot the performance relative to the number of training objects used, as shown in Figure 10, keeping the total number of iterations fixed at 10K. Surprisingly, fine-tuning on just one object already provides significant performance improvements. Additionally, the object was randomly selected from Objaverse. We tested six different objects, all of which yielded similar results. The results are shown in Figure 11. Notably, even simple shapes like an untextured hemisphere can enhance the 3D understanding of the ViTs in these varied tasks.

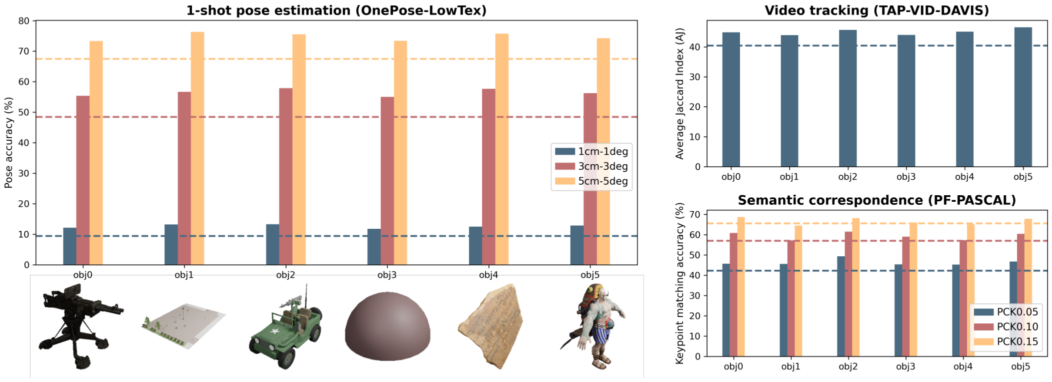

Figure 11: **Finetuning with different objects.** All results are tested with finetuned DINOv2. Dashed lines indicate the performances of the original pretrained model. The feature finetuning method is effective with as few as one single object. It also shows insensitivity to the specific choice of the object, even if the object has limited textures or is uncommon in daily life.

Convergence Within a Few Iterations Figure 12 plots the performance of downstream tasks versus the number of training iterations on a single object. Interestingly, our experiments reveal that training with just a single multi-view pair of one object for a single iteration significantly boosts the model's 3D equivariance, as shown by the sharp improvement at the first elbow of Figure 12. This finding is remarkable, indicating that fine-tuning for 3D correspondence in vision transformers is highly efficient in capturing essential 3D spatial relationships with minimal data. Even with such a minimal training setup, the model effectively learns the desired 3D properties, substantially improving performance across tasks without requiring extensive training or large datasets.

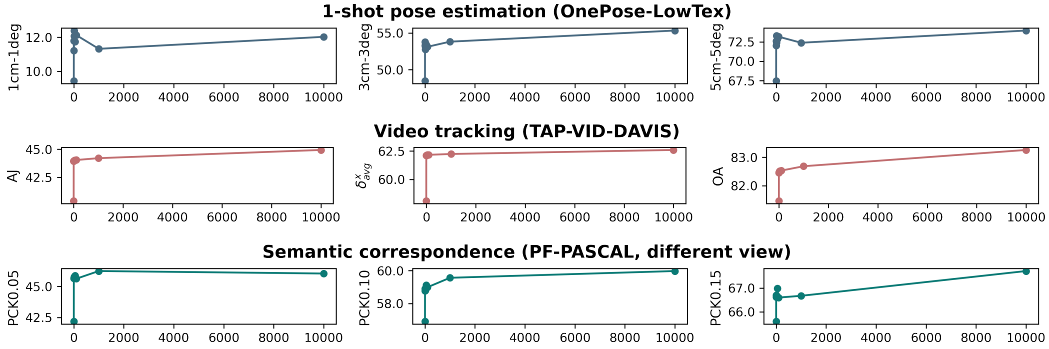

Figure 12: **Finetuned DINOv2 performances *w.r.t.* #training iterations.** We evaluate the performances of the DINOv2 model finetuned with *only one object* over 0, 1, 5, 10, 20, 50, 100, 1000, 10000 training iterations.

4 DESIGN CHOICES FOR FINETUNING

In this section, we ablate and verify the design choices of our finetuning strategy and share some findings. We use the best DINOv2 *base* model for all our ablations.

4.1 ADDITIONAL CONVOLUTION LAYER HEAD

We append a single convolution layer to the original model architecture and find that gives surprisingly good performance. Adding a single convolutional layer to the finetuning architecture was motivated by the need to improve the resolution and consistency of the dense feature maps produced by Vision Transformer (ViT) models. The typical ViT models process images as low-resolution patches, and while global attention mechanisms facilitate communication between patches, they are not optimized for generating dense per-pixel features during interpolation. By incorporating a convolutional layer with a kernel size of 3 and a stride of 1, we can explicitly exchange information between neighboring patches, allowing the model to generate more accurate and high-resolution feature maps before interpolation. We ablated the number of convolutional layers and table 1 shows that one conv layer gives the best performance.

ViT models	**OnePose-LowTex**			**TAP-VID-DAVIS**			**PF-PASCAL (Diff. View)**		
	1cm-1deg	3cm-3deg	5cm-5deg	AJ	δ^x_{avg}	OA	PCK0.05	PCK0.10	PCK0.15
DINOv2-FT (Conv 0)	11.69	53.85	72.83	44.50	60.79	84.08	44.82	57.14	65.26
DINOv2-FT (Conv 1)	**13.58**	**58.03**	**77.35**	46.85	**63.84**	**84.15**	**47.25**	**60.76**	**67.57**
DINOv2-FT (Conv 2)	13.12	56.14	75.45	**47.42**	63.25	84.12	46.32	58.05	64.90
DINOv2-FT (Conv 3)	12.15	53.63	74.46	46.84	62.14	82.90	41.60	53.97	60.22

Table 1: **Ablation on the number of appended conv layers.**

4.2 TRAINING DATA

MVImgNet v.s. Objaverse Our results indicate that finetuning on MVImgNet is slightly worse compared to finetuning on Objaverse, likely due to the denser correspondences provided by Objaverse. Both datasets provide a similar object-centric multi-view setup. Although Objaverse is a synthetic dataset and MVImgNet consists of real-world captures, large foundation models tend to be largely agnostic to the distinction between simulated and real images.

Object-centric datasets v.s. scene-centric datasets An interesting result, as shown in Table 2, is that finetuning on scene-centric datasets (*e.g.* RealEstate10K Zhou et al. (2018), Spaces Flynn et al. (2019), and LLFF Mildenhall et al. (2019), which contain diverse real-world scenes with complex backgrounds, does not necessarily improve the performance but sometimes make it worse (*e.g.* PF-PASCAL). This may indicate that 3D objects themselves have already encoded enough 3D spatial reasoning information. And scene-centric dataset does include much more background clutter that may distract the network, leading to less accurate feature representations.

ViT models	**OnePose-LowTex**			**TAP-VID-DAVIS**			**PF-PASCAL (Diff. View)**		
	1cm-1deg	3cm-3deg	5cm-5deg	AJ	δ^x_{avg}	OA	PCK0.05	PCK0.10	PCK0.15
DINOv2-FT (Objaverse)	13.58	58.03	**77.35**	46.85	**63.84**	**84.15**	**47.25**	**60.76**	**67.57**
DINOv2-FT (MVImgNet)	13.65	56.98	74.61	41.53	58.89	82.67	45.13	57.93	65.40
DINOv2-FT (Scene-Centric)	**15.95**	**60.79**	76.35	**47.36**	63.07	80.27	41.73	52.33	60.33

Table 2: **Ablation on the dataset used for finetuning.** Synthetic object-centric dataset, i.e., Objaverse delivers the best overall performance.

ViT models	**OnePose-LowTex**			**TAP-VID-DAVIS**			**PF-PASCAL (Diff. View)**		
	1cm-1deg	3cm-3deg	5cm-5deg	AJ	δ^x_{avg}	OA	PCK0.05	PCK0.10	PCK0.15
DINOv2-FT (SmoothAP)	**13.58**	**58.03**	**77.35**	**46.85**	**63.84**	**84.15**	**47.25**	**60.76**	**67.57**
DINOv2-FT (Contrastive)	13.28	55.57	75.68	43.79	62.20	81.84	46.70	58.08	66.21
DINOv2-FT (DiffProc)	12.92	55.00	74.86	43.60	61.32	82.74	43.89	57.22	64.66

Table 3: **Ablation on the loss function used.** SmoothAP delivers the best overall performance.

4.3 LOSS FUNCTIONS

We start with naive contrastive loss and found that it does not perform as well. This is because contrastive loss does not directly optimize for the correspondence but optimize for pushes negative sample larger than an arbitrary margin. In contrast, SmoothAP optimizes a ranking loss directly, leading to an improved average precision for feature retrieval between corresponding pixels. We also experimented with the differentiable nearest neighbor and Procrustes alignment loss from Li et al. (2022), but these did not outperform SmoothAP. Detailed comparisons are given in Table 3.

5 RELATED WORKS

Vision Transformers Dosovitskiy (2020) (ViTs) have made significant strides in image understanding by employing self-attention mechanisms to capture global contextual information, outperforming traditional convolutional neural networks (CNNs) in tasks such as image classification and object detection. However, despite their success in 2D applications, adapting these models to grasp 3D spatial relationships remains a challenging and relatively unexplored area.

There is growing interest in assessing the 3D comprehension of vision models. While some studies have investigated how well generative models capture geometric information from a single image Bhattad et al. (2024); Du et al. (2023); Sarkar et al. (2024), these efforts are generally specific to generative models, limiting their applicability to broader vision tasks. More closely aligned with our work is El Banani et al. (2024), which evaluated the 3D awareness of visual foundation models through task-specific probes and zero-shot inference using frozen features. In contrast, we delve deeper into this topic and introduce a simple yet effective method for finetuning 3D awareness in ViTs.

Several researchers have also explored applying large-scale models to 3D tasks. For instance, some approaches utilize features from pre-trained models for tasks such as correspondence matching Zhang et al. (2023); Cheng et al. (2024) and pose estimation Örnek et al. (2023). ImageNet3D Ma et al. (2024) studies how global tokens from ViT vary across different views and benefit pose estimation. Their work explores a complementary area to ours, as they focus on view-dependent global features, whereas we emphasize dense, pixel-level features that are invariant to viewpoint changes. ImageNet3D's top-down approach to pose estimation relies on classifying different poses using pretrained features with an added domain-specific linear layer. However, as our paper focuses on general-purpose ViT features, it is hard to apply their method across different dataset domains like OnePose-LowTex and YCB-Video. For general unseen tasks and datasets, we argue finding correspondences—or equivalently, learning equivariant representations, is a better approach. Features that vary across viewpoints are unsuitable for general-purpose settings since we neither know nor can control their variation.

Recent works, such as FiT Yue et al. (2024) and DVT Yang et al. (2024), attempt to finetune pretrained features. FiT lifts 2D features into 3D space and then projects them back into 2D to enforce 3D consistency. However, this lifting process often results in information loss, leading to suboptimal performance. DVT, on the other hand, implements a denoising process to reduce periodic noise artifacts in images, a method that is orthogonal to our approach. Additionally, DUSt3R Wang et al. (2024) directly predicts 3D coordinates for each 2D pixel, but it lacks a shared consistent feature space and forfeits the rich semantic information provided by large vision models. Furthermore, DUSt3R's correspondences are noisy and only support single-pair outputs, limiting its scalability.

6 LIMITATION AND CONCLUSION

In this work, we systematically evaluated the 3D awareness of large vision models, with a specific focus on their ability to maintain view equivariance. Our comprehensive study demonstrates that current vision transformers, particularly DINOv2, exhibit strong 3D equivariant properties, which significantly correlate with their performance on downstream tasks such as pose estimation, video tracking, and semantic transfer.

Building on these insights, we introduced a simple yet effective finetuning method that enhances the 3D understanding of 2D foundation models. By leveraging multiview correspondences and applying a loss function that enforces feature consistency across views, our approach yields substantial improvements in task performance with minimal computational overhead. Remarkably, even a single iteration of finetuning on a multi-view pair can lead to notable performance gains.

Our findings highlight the importance of 3D equivariance in vision models and provide a practical path to improving 3D understanding in existing models. We believe this work opens up new opportunities for enhancing the 3D capabilities of vision transformers, paving the way for future advancements in image-based 3D scene understanding and generation tasks. All code and resources will be made publicly available to support further research in this direction.

7 STATEMENTS

Ethics Statement. Our method leverages open-sourced simulation data and real data whose data collection process follows strict ethical guidelines. In using these data, we follow the same ethical considerations to protect sensitive information. There is no ethical concerns detected of the proposed method to our knowledge, and we will strive to adhere to ICLR code of conduct for future use of the proposed method.

Reproducibility Statement. We provide extensive details for ease of re-implementation. We strive to ensure our method is reproducible and the findings in this paper are generalizalble. We will release code, results, and scripts for reproduction to promote future research in 3D deep learning.

REFERENCES

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.

Anand Bhattad, Daniel McKee, Derek Hoiem, and David Forsyth. Stylegan knows normal, depth, albedo, and more. *Advances in Neural Information Processing Systems*, 36, 2024.

Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. Smooth-ap: Smoothing the path towards large-scale image retrieval. In *European conference on computer vision*, pp. 677–694. Springer, 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*, pp. 9650–9660, 2021.

Xinle Cheng, Congyue Deng, Adam Harley, Yixin Zhu, and Leonidas Guibas. Zero-shot image feature consensus with deep functional maps. *arXiv preprint arXiv:2403.12038*, 2024.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13142–13153, 2023.

James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Xiaodan Du, Nicholas Kolkin, Greg Shakhnarovich, and Anand Bhattad. Generative models: What do they know? do they know things? let's find out! *arXiv preprint arXiv:2311.17137*, 2023.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21795–21806, 2024.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338, 2010.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2367–2376, 2019.

Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1711–1725, 2017.

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*, pp. 16000–16009, 2022a.

Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without cad models. *Advances in Neural Information Processing Systems*, 35:35103–35115, 2022b.

Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. *arXiv preprint arXiv:2305.15581*, 2023.

Tomáš Hodaň, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiří Matas. Bop challenge 2020 on 6d object localization. In *European Conference on Computer Vision*, pp. 577–594. Springer, 2020.

Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5885–5894, 2021.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19729–19739, 2023.

Kurt Koffka. *Principles of Gestalt psychology*, volume 44. Routledge, 2013.

Wolfgang Köhler. Gestalt psychology. *Psychologische Forschung*, 31(1):XVIII–XXX, 1967.

Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. *arXiv preprint arXiv:2407.08447*, 2024.

Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022.

Lei Li, Hongbo Fu, and Maks Ovsjanikov. Wsdesc: Weakly supervised 3d local descriptor learning for point cloud registration. *IEEE Transactions on Visualization and Computer Graphics*, 29(7): 3368–3379, 2022.

Wufei Ma, Guanning Zeng, Guofeng Zhang, Qihao Liu, Letian Zhang, Adam Kortylewski, Yaoyao Liu, and Alan Yuille. Imagenet3d: Towards general-purpose object-level 3d understanding. *arXiv preprint arXiv:2406.09613*, 2024.

Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (ToG)*, 38(4):1–14, 2019.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomas Hodan. Foundpose: Unseen object pose estimation with foundation features. *arXiv preprint arXiv:2311.18809*, 2023.

Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5706–5715, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David A Forsyth, and Anand Bhattad. Shadows don't lie and lines can't bend! generative models don't know projective geometry... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28140–28149, 2024.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pp. 746–760. Springer, 2012.

Emma EM Stewart, Frieder T Hartmann, Yaniv Morgenstern, Katherine R Storrs, Guido Maiello, and Roland W Fleming. Mental object rotation based on two-dimensional visual representations. *Current Biology*, 32(21):R1224–R1225, 2022.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *European conference on computer vision*, pp. 516–533. Springer, 2022.

Narek Tumanyan, Assaf Singer, Shai Bagon, and Tali Dekel. Dino-tracker: Taming dino for self-supervised point tracking in a single video. *arXiv preprint arXiv:2403.14548*, 2024.

Thomas Vetter, Anya Hurlbert, and Tomaso Poggio. View-based models of 3d object recognition: invariance to imaging transformations. *Cerebral Cortex*, 5(3):261–269, 1995.

Qianxu Wang, Haotong Zhang, Congyue Deng, Yang You, Hao Dong, Yixin Zhu, and Leonidas Guibas. Sparsedff: Sparse-view feature distillation for one-shot dexterous manipulation. *arXiv preprint arXiv:2310.16838*, 2023.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

Kevin D Wilson and Martha J Farah. When does the visual system use viewpoint-invariant representations during recognition? *Cognitive Brain Research*, 16(3):399–415, 2003.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

Jiawei Yang, Katie Z Luo, Jiefeng Li, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Denoising vision transformers. *arXiv preprint arXiv:2401.02957*, 2024.

Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9150–9161, 2023.

Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2d feature representations by 3d-aware fine-tuning. *arXiv preprint arXiv:2407.20229*, 2024.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *arXiv preprint arXiv:2305.15347*, 2023.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.