Probing the 3D Awareness of Visual Foundation Models

Mohamed El Banani[1] Amit Raj[2] Kevis-Kokitsi Maninis[2] Abhishek Kar[2] Yuanzhen Li[2]
Michael Rubinstein[2] Deqing Sun[2] Leonidas Guibas[2] Justin Johnson[1] Varun Jampani [2*]
[1] University of Michigan [2] Google Research

Abstract

Recent advances in large-scale pretraining have yielded visual foundation models with strong capabilities. Not only can recent models generalize to arbitrary images for their training task, their intermediate representations are useful for other visual tasks such as detection and segmentation. Given that such models can classify, delineate, and localize objects in 2D, we ask whether they also represent their 3D structure? In this work, we analyze the 3D awareness of visual foundation models. We posit that 3D awareness implies that representations (1) encode the 3D structure of the scene and (2) consistently represent the surface across views. We conduct a series of experiments using task-specific probes and zero-shot inference procedures on frozen features. Our experiments reveal several limitations of the current models. Our code and analysis can be found at https://github.com/mbanani/probe3d.

1. Introduction

Large-scale pretraining on image datasets has yielded visual foundation models with impressive generalization capabilities. Such models can classify [46, 65], segment [36], and generate [10, 69, 70] arbitrary images. Furthermore, the dense representations learned by such models extend beyond their training tasks and exhibit strong zero-shot capabilities in other tasks such as segmentation [56, 95] and part discovery [1, 27]. This suggests that models are learning strong image representations, but how well do they represent the 3D world that images depict?

Recent work suggests that visual foundation models are useful for some 3D tasks despite being trained with 2D data. For instance, models implicitly represent depth and surface normals when generating images of scenes and faces [6, 12]. Furthermore, the intermediate representations of self-supervised and generative models can be used to estimate semantic correspondence [1, 27, 30, 83, 99] and object pose [25]. However, when reconstructing 3D objects, they generate artifacts that suggest a lack of 3D consistency [50]; *e.g.*, animals with multiple faces. Therefore, it remains unclear how those modes represent or understand the 3D world.

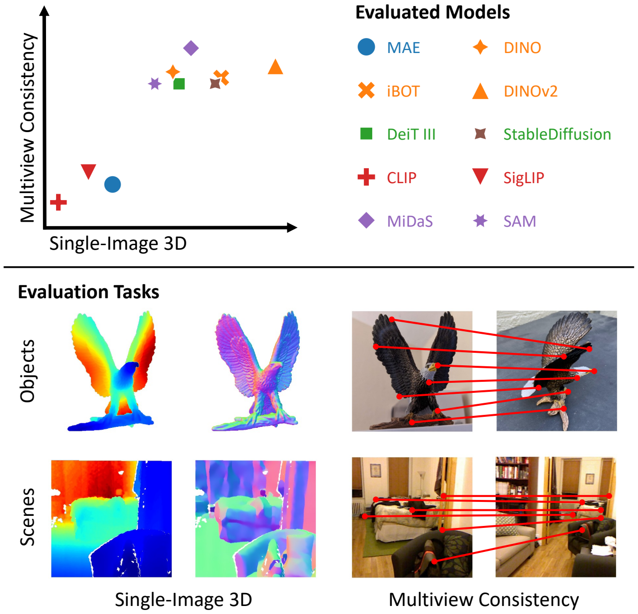

Figure 1. **Are current visual foundation models 3D aware?** We probe the 3D awareness of the learned representations by evaluating their ability to encode the 3D structure of the visible surface and their consistency across views.

The goal of this paper is to study the 3D awareness of visual foundation models. Previous benchmarks evaluate visual models on semantic tasks [24, 26, 87], but their 3D understanding remains understudied. Representations can vary from having no 3D awareness (*e.g.*, class label or bag of words) to accurately representing the 3D geometry of the scene (*e.g.*, 3D surface map or mesh). We posit that 3D awareness can be evaluated through two distinct capabilities: *single-view surface reconstruction* and *multiview consistency*. If a model is 3D aware, we expect that its representations would encode the geometry of the surfaces visible in the image; *i.e.*, how far is each point in the scene? what is the orientation of the surface? Moreover, the representations should be consistent for different views of the scene; allowing them to establish accurate correspondence.

[*] Current affiliation is Stability AI.

To this end, we conduct an empirical analysis of the 3D awareness of visual foundation models. Our analysis considers a range of large-scale pretrained models that have exhibited strong generalization, regardless of their pretraining objective. We evaluate the models on their ability to estimate 3D quantities that match the aforementioned capabilities: depth, surface normals, and 3D correspondence. Furthermore, we evaluate those capabilities at both the scene level [13, 78] and for individual objects [32, 57] to provide further differentiation. We show the models and tasks considered in Fig. 1. Since we are interested in what the models represent, we probe the frozen representations through task-specific probes or zero-shot inference methods. This allows us to evaluate the models' representations, rather than the transferability of their pretrained weights.

Our experiments reveal a large variation in the 3D awareness of the models. We present the aggregated ratings (higher is better) of different models in single-image and multiview tasks in Fig. 1. We find that recent self-supervised models such as DINOv2 [60] learn representations that encode depth and surface normals, with StableDiffusion [69] being a close second. Meanwhile, the training in vision language for models such as CLIP [65] exhibits very poor performance despite its impressive semantic generalization capabilities. At their best, some of the probes achieve a performance close to that of state-of-the-art models despite being pretrained with a very different objective. Meanwhile, we find that the models struggle with multiview consistency. Although most models can accurately match objects and scenes with small viewpoint changes, they perform very poorly at large viewpoint variations. Our analysis further suggests that consistency across images is semantic in nature; *i.e.*, models accurately match semantic parts but struggle to incorporate the global object pose. We hope that our findings will spark more interest in better understanding the 3D awareness of vision foundation models and contribute to more comprehensive benchmarks of visual representation learning approaches.

2. 3D Aware Visual Representations

We first discuss what we mean by 3D aware visual representations. When we view a scene, we seem to effortlessly understand its 3D structure despite only seeing its 2D projection. Research in developmental psychology and psychophysics suggests that our perception encodes surface properties such as depth and orientation [39, 79]. Research on mental imagery has suggested that our internal representations of objects encode their 3D shape and are subject to 3D constraints [76]. Inspired by this work, we posit that 3D aware representations encode basic 3D properties of the surface as distances and orintations. Beyond a single image, 3D aware representations are consistent across views of the same object or scene, as they are projections of the same underlying 3D geometry.

Representations in computer vision have varied a lot in how well they represented the 3D shapes of objects. Early representations such as 2.5D sketches [55] and generalized cylinders [7, 8] explicitly depicted the 3D geometry of the obejcts and their spatial relationships. Recent advances have deviated from explicit modeling and instead rely on the representation of visual information as dense feature grids [28] or sets of tokens [15]. While 3D awareness of early representations was obvious, it remains unclear what the learned representations encode or how 3D aware they are. Popular interpretability mechanisms such as GradCAM [74] are not helpful here, as they tell us which components of the image led to a specific inference, not what information was represented by the network.

We propose evaluating the 3D awareness of visual models by probing them on two capabilities: single-view 3D and multiview consistency. We take inspiration from the work on human perception [38, 75, 79] and evaluate models on how well they encode basic 3D properties and how 3D consistent they are. For a single image, we expect a 3D aware model to accurately represent the visible surface and encode properties such as depth and surface orientation. When given multiple images of the same object or scene, we expect a 3D aware representation to capture the relationships between the images and provide accurate correspondence. Although these two capabilities are not exhaustive, they capture two fundamental aspects of 3D understanding. Furthermore, they can be directly mapped to three well-studied problems in computer vision, namely, estimating monocular depth, surface normals, and correspondence.

3. Experimental Setup

The goal of our experiments is to evaluate the 3D awareness of visual foundation models: *i.e.*, large-scale pretrained models that are proposed as general backbones for a wide variety of downstream tasks or applications. Specifically, we hope to answer the following questions:

1. Do models learn to represent the visible surface?
2. Are the representations consistent across views?
3. How does the training objective impact 3D awareness?

Models. We primarily focus our experiments on vision transformers that were proposed as general purpose backbones or that exhibit strong generalization performance across tasks or domains. Moreover, we are interested in evaluating models that were trained with different supervisory signals. First, we consider three forms of supervision that commonly serve as pretraining tasks: classification [86], language supervision [31, 65], and self-supervision [9, 29, 60, 102]. Recent work has also shown that text-conditioned image generation [69] can learn strong representations and provide strong backbones for other vision tasks [45, 95, 101]. We also consider two forms of dense supervision that have recently

Table 1. **Evaluated Visual Models.** We consider a range of visual models spanning several forms of supervision. We evaluate publicly available checkpoints and choose checkpoints of comparable model and training size whenever possible.

Model	Architecture	Supervision	Dataset
DeIT III [86]	ViT-B/16	Classification	ImageNet-22k
MAE [29]	ViT-B/16	SSL	ImageNet-1k
iBOT [102]	ViT-B/16	SSL	ImageNet-1k
DINO [9]	ViT-B/16	SSL	ImageNet-1k
DINO v2 [60]	ViT-B/14	SSL	LVD-142M
CLIP [65]	ViT-B/16	VLM	WIT-400M
SigLIP [97]	ViT-B/16	VLM	WebLI
StableDiffusion [69]	UNet	Generation	LAION
SAM [36]	ViT-B/16	Segmentation	SA-1B
MiDaS [67]	ViT-L/16	Depth	MIX-6

been scaled up: depth estimation [67, 68] and class-agnostic segmentation [36]. While there models have not been used as general purpose backbones yet, they exhibit impressive generalize to a wide range of domains and provide an interesting point of comparison. We present an overview of the models considered in Tab. 1, and more details can be found in App. A.1.

One challenge is how to fairly compare models that have different data and compute requirements. This challenge is further amplified by considering the scale used to achieve the strong performance displayed by such models. Furthermore, the data used to train many of these models is private [60, 65] and even replicating the data collection and curation process requires extensive resources as shown by Xu et al. [94]. Beyond data scale and curation, models have different data requirements that range from class labels [86], captions [65], masks [36], or even simple curation [60]. As a result, it is unclear which dataset would provide a fair comparison. We make a pragmatic choice of relying on publicly available checkpoints and selecting checkpoints of comparable architectures and training scale to provide some fair comparison. We provide additional comparisons in App. B and discuss the impact of this on our results in App. C.

Another important question is how to evaluate those properties. One common approach is transfer learning, where the pretrained model is fine-tuned using task-specific supervision. This is often a good practical choice, as it results in strong downstream performance. However, it is not suitable for our analysis, as good fine-tuning performance may indicate two different things: the model has good 3D awareness or the model weights are a good initialization for other tasks [26]. Furthermore, fine-tuning specializes the models by sacrificing its generality [42]. Instead, we probe the frozen features with trainable probes and zero-shot inference methods that do not change model weights or significantly alter model capacity. This allows us to evaluate pretrained representations of models with the assumption that the same model may be used for a wide range of tasks.

3.1. Single Image Surface Reconstruction

In this section, we analyze how well the models represent the visible surface in the image. We consider two tasks for single-view 3D understanding: depth estimation and surface normal estimation. Those tasks are well established in computer vision and are commonly studied in human perception and development [79]. Although depth and surface normals are closely related quantities, they are different prediction tasks as they rely on different visual cues, as discussed by Koenderink and Van Doorn [39] and Fouhey [22]. We briefly outline our evaluation setup below, and refer the reader to App. A and our code release for more specific details.

Monocular depth estimation is the task of predicting the depth for each pixel in the image. Although early work framed the task as regression [17], recent work has shown that the use of binned prediction results in better performance Bhat et al. [4]. We follow the AdaBins [4] formulation and train dense probes using their proposed losses. We report the root-mean-squared prediction error for depth estimation as well as recall at different threshold rations, similar to Eigen et al. [17].

We find that estimating the depth for object-centric datasets is particularly challenged by scale ambiguity. While scale ambiguity affects both objects and scene, we find that models trained to estimate metric depth on objects end up focus on predicting the object's mean depth without capturing any details. As a result, we use a scale-invariant formulation for objects by normalizing their depth between 0 and 1.

Surface normal estimation is the task of predicting the orientation of the surface at every pixel. We adopt the setup of Bae et al. [2], which utilizes an uncertainty-aware angular loss. Similarly to Fouhey et al. [23], we report the root-mean square angular prediction error as well as the precentage recall at different angular thresholds.

Probe. We use a dense multiscale probe similar to the DPT decoder [68]. This deviates from the common choice of linear probing commonly used in self-supervised model benchmarking [41]. Linear probing is useful for semantic tasks since linear separability of classes is a desired and expected property. However, it is unclear why we should require the encoding of 3D properties to be linear. Furthermore, the model may represent such properties at different, or multiple, locations within the network. Hence, instead of training a linear probe on a specific linear, we use a multiscale dense probe to map the features from multiple layers to either depth or surface normals.

Optimization. We train the probes for 10 epochs using the AdamW [35, 52] optimizer with a linear warm and cosine decay learning rate scheduler. While longer training further improves performance, trends stabilize after 5 training epochs due to the relatively small capacity of the probe.

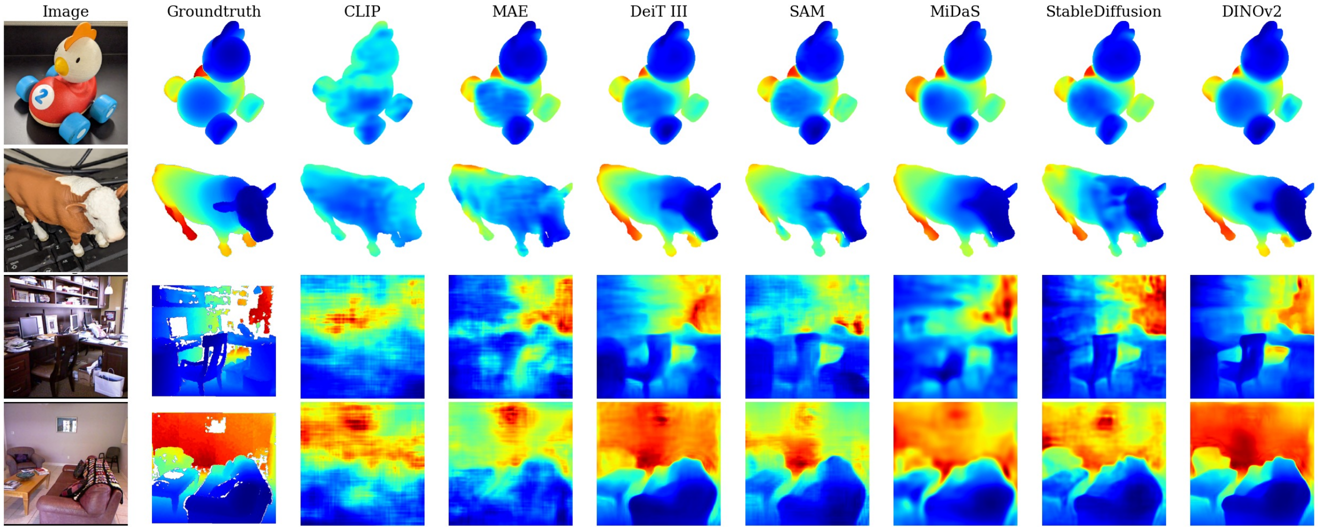

Figure 2. **Depth Estimation Results.** While pretrained representations exhibit large variation in their ability to represent depth, their performance is consistent on objects and scenes. CLIP and MAE features do not encode depth and appear to instead capture rough priors such as ”floor pixels are close”. Most models appear to capture the rough structure of the scene and vary in the degree to which they capture details. DINOv2 performs best and accurately captures fine details; *e.g*., cow's ear, desk chair, and coffee table.

Datasets. We evaluate the performance on both scenes and objects. We use the NYUv2 dataset [78] to evaluate scene-level performance as it is a common benchmark for indoor scene understanding. We evaluate object-level performance using the NAVI dataset [32] which includes a set of object instances in a wide range of scenes and orientations. Both datasets provide aligned depth maps. For surface normals, we use the annotations generated by Ladickỳ et al. [43] and generate the surface normal annotations for NAVI.

Results. We evaluate all models and report the performance in App. B due to space limitations. We focus here on qualitative results and performance trends, and analyze them through a series of questions:

Do models learn to represent depth? We observe that the ability of the models to encode depth is highly variable. This can be clearly seen in Fig. 2 where DINOv2 and StableDiffusion predict accurate and detailed depth maps that capture the cow's ear and chair legs, while CLIP and MAE generate blurry and inaccurate estimates. It is worth noting that the models compared are all highly performant models that are often used within as backbones for downstreams taks. The disparity seen highlights the importance of considering a wider range of tasks for benchmarking such models, as well as the utility of 3D awareness as a domain for such benchmarking.

Do models learn to represent surface normals? Surface normal probe results reveal similar trends to depth estimation, with some models achieving very high performance, while others struggle to capture any information beyond the coarse priors such as “floor pixels point up.” The reliance on priors becomes more clear when comparing predictions for objects and scenes since objects have fewer priors due to the large pose variation. This is useful when analyzing the qualitative results of CLIP, which may appear blurry but correct for scenes, but are clearly inaccurate for objects. However, the best-performing model, DINOv2, achieves an impressive performance that is competitive with state-of-the-art models.

How is performance correlated across both tasks? We observe that the performance of models is strongly correlated across domains and tasks as shown in Fig. 4. This supports our experimental design choices as it suggests that we are measuring a single capability using different methods. Furthermore, the consistent performance across indoor scenes and objects suggests that such models are learning to represent some information about the visible surface without any task-specific supervision. Although recent work has focused on the ability of generative models to learn this information [6, 12], we find that it is not unique to such models trained with classification or discriminative self-supervision achieving comparable performance.

We note that, while depth and surface normal performance are well correlated at the model level, the correlation is far weaker when considering performance at the image or pixel level. We find that model performance is not consistent at the image or patch level; *e.g*., we find that the correlation between errors made by DINOv2 on NYU is 0.37 when aggregating at the image level, and 0.13 when considering pixel-level errors. Hence, while the underlying ability to represent the surface is shared, surface normals and depth estimation rely on different visual cues[22, 40, 59] resulting in model errors being weakly correlated.

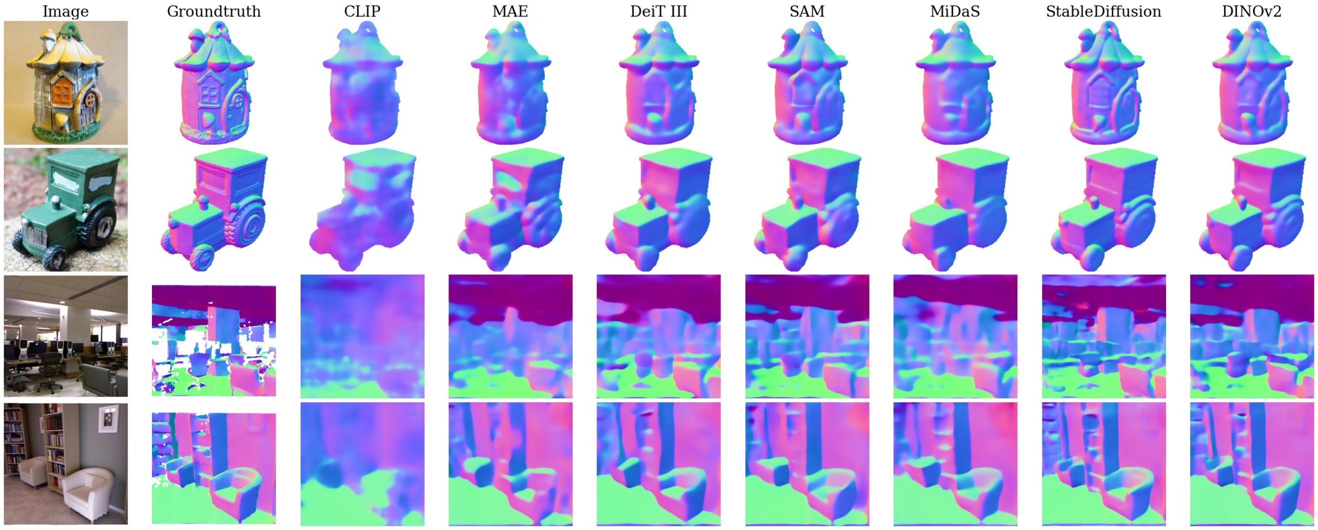

Figure 3. **Surface Normal Qualitative Examples.** With the exception of CLIP, models can capture the rough orientation of object and scene surfaces; *e.g.*, floors, walls, ceilings. The main distinction seems to be in how well they capture finer details. Similarly to depth results, we find that DINOv2 and StableDiffusion perform best and can capture fine details such as the edges of the toy car and the white seat. Surprisingly, we find that SAM's predictions are not as detailed despite its ability to predict accurate segmentation boundaries.

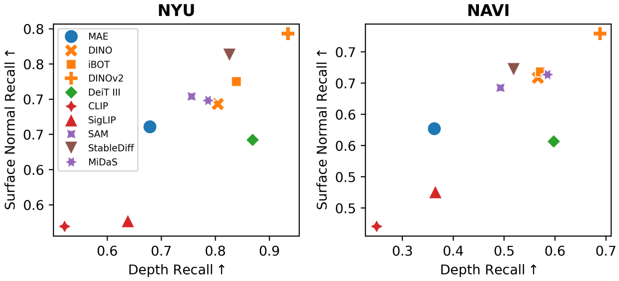

Figure 4. **Single view performance correlation.** Depth and surface normal performance is highly correlated across domains.

What is the impact of the training objective? We observe that discriminative self-supervised models perform best across both tasks and domains. This is surprising since it is unclear why the self-distillation and instance descrimination losses used to train such models would encourage this behavior. Consistent with other work [6, 12], we find that StableDiffusion also captures surface properties well. Interestingly, models trained with dense supervision, even depth supervision, perform worse than self-supervised and text-conditioned generation, and perform on par with classification-trained models. Finally, language-supervised models appear to perform poorly despite their common utility as backbones for a variety of tasks. This could be related to previous findings that vision language models struggle with spatial relations and compositionality [44, 48, 81].

Overall, our experiments suggest that most visual models suggest that most visual foundation models end up learning representations that encode properties of the visual surface despite being trained with just image data.

3.2. Multiview Consistency

We previously evaluated the models' ability to represent the visible surfaces. Although this is important for 3D understanding, the evaluation is limited to a single image. As discussed previously, 3D awareness also implies consistency of representations across multiple views. We evaluate this using correspondence estimation, where the goal is to identify image patches across views that depict the same 3D point. This capability is important because it would allow the model to correctly aggregate information across views, which is central to reconstruction and localization pipelines.

Geometric correspondence estimation. Given two views of the same object or scene, identify pixels across views that depict the same point in 3D space. Rather than training a probe, we directly compute correspondence between the dense feature maps extracted from each image as this allows us to directly evaluate the consistency of the representations. This inference procedure is derived from keypoint-free correspondence estimation pipelines [18, 19, 82] and is similar to recent approaches to assess feature quality [1, 83, 99]

Datasets. We consider both scenes and objects. For scenes, we evaluate our model on the Paired ScanNet [13] split proposed by Sarlin et al. [72]. For objects, we sample view pairs from the NAVI wild set which depict the same object instances in different environments. We sample views that have a maximum rotation of 120 degrees to ensure that there exists a mutually visible surface. We also evaluate performance on the SPair dataset [57] which provides keypoint-labeled images allowing us to analyze the models' performance on a closely related task: semantic correspondence estimation.

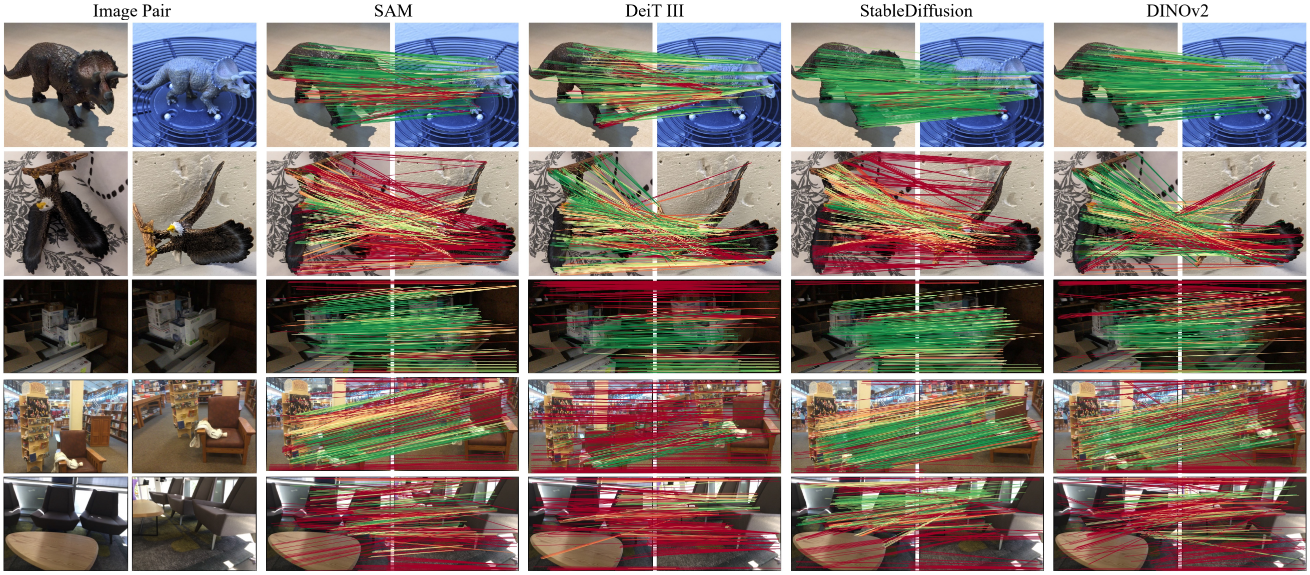

Figure 5. **Correspondence Estimation Qualitative Results.** We observe that models can estimate accurate correspondence for small viewpoint changes, but struggle with large viewpoint changes. This is true even if the change is an in-plane rotation as shown with the eagle. This pattern is consistent for both objects and scenes, although performance is not well correlated: SAM and StableDiffusion perform better for scenes, while DeiT and DINOv2 are more consistent for objects. Correspondence color-coded for accuracy.

Evaluation. We report the correspondence recall; *i.e.*, the percentage of correspondence that falls within some defined distance. Correspondence error is often computed in pixels to account for the large variation in depth; *e.g.*, a prediction off by 1 pixel can be a few millimeters on a near-by surface or several meters for outdoor scenes. This choice is less suitable for objects, since they do not have the same large variation depth. Object can also suffer from self-occlusion and repeated parts, which makes a pixel-wise threshold potentially errenous. Therefore, we use a metric threshold for objects. Since layer selection can greatly affect performance [87], we evaluated model performance at four different intermediate points. Finally, we find that model performance varies greatly depending on the viewpoint differnce between the view pairs, as we discuss next. As a result, we bin the performance depending on the magnitude of the transformation between the view pairs. For more details on the evaluation setup, we refer the reader to App. A.

We evaluate all models on the three datasets and report the results in App. B. We present qualitative results and performance trends in Fig. 5 and Fig. 6.

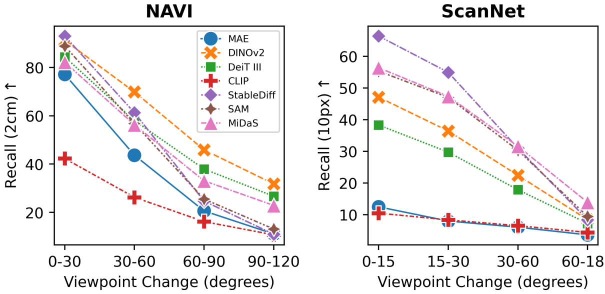

Figure 6. While all models experience performance drops with larger viewpoint changes, some experience sharper drops suggesting a lack of 3D awareness.

Are the representations 3D consistent? While models can estimate accurate correspondence between objects for small viewpoint changes, the performance quickly deteriorates for larger viewpoint changes, as seen in Fig. 6. Although we expect performance to be lower for larger viewpoint changes as they are more difficult, the rate of deterioration is interesting. Specifically, StableDiffusion and SAM experience very sharp drops from being among the top performers for the smallest viewpoint changes to being the worst models for the larger viewpoint changes. This can be clearly seen in Fig. 5 where both models predict accurate dense correspondence for the dinosaur in the top row, where the viewpoint variation is minimal, but perform very poorly for the rotated eagle views. This rapid deterioration is not universal, as shown by the wide baseline performance of DINOv2 and DeiT.

We observe similar trends for indoor scenes where the models predict accurate correspondence when viewing the scene from a very similar vantage point, but struggle with even small viewpoint changes as seen in the last two rows of Fig. 5. Although DINOv2 performs better than the other models, the absolute performances for all models are very low for wide baseline correspondence estimation. In general, our results suggest that current models are not 3D consistent despite encoding surface properties as shown in Sec. 3.1.

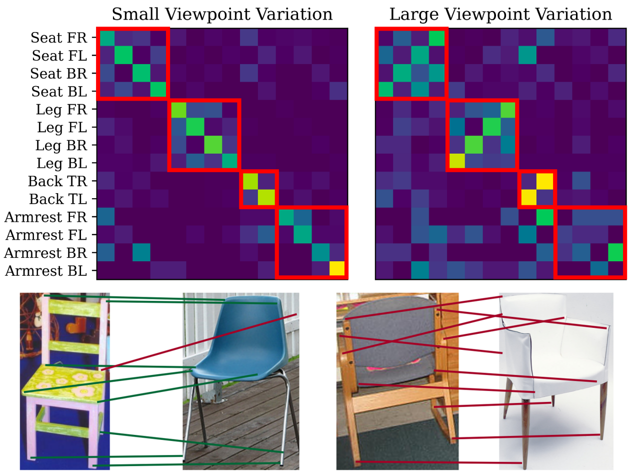

Figure 7. **Semantic Correspondence.** StableDiffusion represents semantics well, but lack 3D consistency. This results in accurate correspondence for objects viewed from similar angles and systematic errors when viewing objects from different viewpoints.

Semantic vs. Geometric Correspondence. Recent work has shown that self-supervised and generative models excel at estimating semantic correspondence [1, 83, 99]. Semantic correspondence [3] generalizes the correspondence problem from matching the same points across views of the same object to matching similar semantic parts across different instances of the same class; *e.g*., matching a dog's left ear in images of two different dogs. At first glance, this seems to contradict our results, since semantic correspondence appears to capture both 3D structure and semantics.

Semantic correspondence is commonly evaluated using keypoint recall. This evaluation makes the model's performance susceptible to semantic biases and priors in the data. Keypoints are often selected to be unique and easily identifiable; *e.g*., beaks and tails. Although some keypoints (*e.g*., eyes and knees) are repeated, they often appear in consistent spatial arrangements due to photographer bias.

We illustrate the disparity between semantic and geometric correspondence by evaluating StableDiffusion on SPair-71k chairs in Fig. 7. We evaluate performance using keypoint confusion rather than recall. We do this by matching the closest keypoint to the predicted correspondence location and plotting the confusion matrix. This is only computed for keypoints with a true match. While StableDiffusion estimates accurate correspondence for small viewpoint changes, it exhibits interesting error patterns for large viewpoint changes. Errors seem restricted to semantically related classes (*e.g*., seat corners, and chair legs). Furthermore, the qualitative results suggest that the representation captures a combination of semantics and 2D location: *i.e*., the chair leg on the right. We suspect that this observation is related to the Janus problem observed in diffusion-based 3D reconstruction, since the same ear can be repurposed for two different faces.

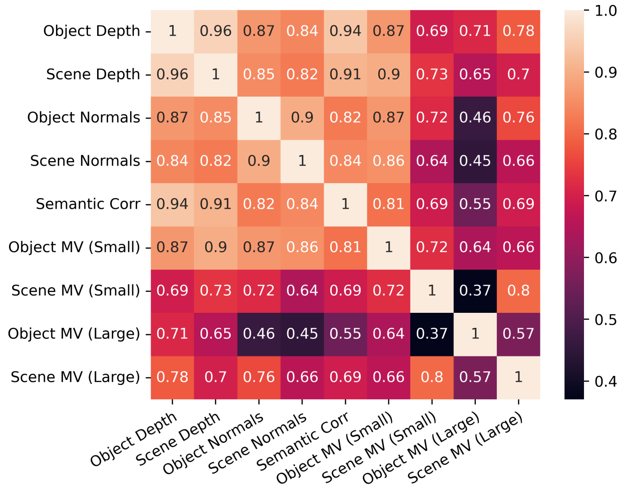

Figure 8. **Cross-task performance correlation.** Performance on single view tasks is strongly correlated with itself as well as semantic correspondence, but we see a drop in correlation performance of scene-level correspondence estimation and correspodence estimation with large viewpoint variation.

3.3. Analysis

One important question is how correlated are different tasks; *i.e*., if a model's representations accurately represent depth, how likely is it that they are also useful for correspondence? To address this question, we compute the correlations between the models' aggregated performance across multiple tasks. We are particularly interested in understanding the relationship between training objectives and 3D awareness. We note that while we highlighted specific models in our analysis, we evaluated a much larger set of model variants and computed the cross-task performance correlations on the full set. See App. B for the complete set of results.

We compute the Pearson correlation between all pairs of tasks as shown in Fig. 8. For single-view 3D, we report recall for depth and surface normal estimation on objects and scene. We also report recall for correspondence estimation and separate the performance based on the amount of viewpoint variation by considering the smallest and largest viewpoint bins for NAVI and ScanNet. Finally, we also report the aggregated performance for semantic correspondence estimation.

We find that performance on all single view tasks is strongly correlated with correlation coefficients larger than 0.82. On the other hand, the correlation across multiview tasks is much lower, as shown by the values on the bottom right corner of the correlation matrix. Interestingly, semantic correspondence performance is more strongly correlated with single-view tasks than it is with multiview tasks despite having a similar evaluation procedure to the latter. This further supports our claim that semantic correspondence is not a good measure of 3D consistency.

4. Related Work

Our work is broadly related to other efforts to understand the representations learned by vision models and to use them for 3D vision tasks. Since the recent revival of deep learning, there has been a lot of work on understanding how and what these models learn with a focus on classification models. Early work focused on analyzing what those models could be used for [11, 26, 41] and providing some interpretability into what they were learning [74]. Our work is inspired by recent efforts to benchmark the semantic and localization capabilities of visual backbones [24, 26, 44, 48, 85, 87] which we try to extend towards 3D awareness.

Recent work has attempted to evaluate the 3D understanding of vision models. One line of work has explored how well generative models capture single image geometry [6, 12, 16, 71]. Although this line of work shared our goals, their probing techniques are often specific to generative models, making it difficult to extend to other visual models. More closely related to our analysis is the recent work of Zhan et al. [98], who proposed analyzing the 3D understanding of StableDiffusion through a series of binary classification tasks. Instead, we focus on dense probing tasks and multiview consistency, as they are less susceptible to semantic priors, which can confound 3D undersanding, as shown by Tatarchenko et al. [84]. Furthermore, we explore multiview consistency as another facet of 3D awareness.

Another line of work has focus on using large-scale models for 3D tasks. One line of work extracts features from models for correspondence estimation [1, 30, 54, 60, 83, 99] and pose estimation [25, 100]. Others have shown how these models could be fine-tuned for accurate depth estimation with [33, 101] achieving state-of-the-art performance by fine-tuning StableDiffusion. Another line of work combines image generation with 3D representations for text- or image-conditioned 3D reconstruction [62, 88, 93]. While those methods generate impressive 3D shapes, it has been observed that their generations are not 3D consistent and can generate animals with multiple heads (the Janus problem). Recent efforts have shown that fine-tuning with 3D data can improve generation quality [34, 50, 63, 66, 77]. We are inspired by this line of work, but note that it differs in objective from our analysis, as we are interested in understanding 3D awareness in models trained *without* 3D supervision.

5. Discussion

This paper presents an exploratory study of the 3D awareness of visual models; *i.e.*, how well do the representations capture the 3D-ness of the scenes and objects? We posit that 3D awareness implies representations that (1) encode the geometry of the visible surface and (2) are consistent across views. We used trainable probes and zero-shot inference methods to evaluate the frozen features of those models.

Our results show that visual foundation models learn representations that encode the depth and orientation of the visible surface, with vision-language models being the notable exception. We also find that while models can estimate accurate semantic correspondence as well as correspondence across images of a similar viewpoint, they struggle with large viewpoint changes. This indicates a lack of multiview consistency and suggests that models are learning representations that are view-consistent, not 3D consistent. One possibility is that the models are learning view-dependent representations. This could be similar to the theories of shape perception proposed by Koenderink and Van Doorn [37, 38], where shape perception is achieved by a series of view-specific representations connected with an aspect graph. Another possibility is that current models are simply good "image models" and that good discriminative features are sufficient for strong 2.5D understanding. We hope that our findings can simulate more interest in understanding the 3D awareness of visual models and that future work can provide better answers.

Our analysis struggles with several limitations, which we discuss in more detail in App. C. First, we used pretrained checkpoints that were often trained on different datasets and with different compute scales. While this allowed us to explore a broader set of models and tasks, it would be useful to make more targeted and fair comparisons to better understand the impact of training signals. Second, we focused on minimal probing approaches to analyze the pretrained representations. It would be useful to explore other probing techniques, as it remains unclear what is the best way to understand the distributed representations learned by visual models. Finally, our analysis only explored two basic aspects of 3D understanding. However, 3D awareness and understanding are closely related to more complex and higher-order tasks such as perceiving 3D shapes, reasoning about spatial relationships, as well as making predictions about deformation and dynamics.

This work is only a first step towards understanding the 3D awareness of visual models. This is becoming more relevant, as recent image and video generation models have achieved impressive feats of photorealism and temporal consistency. This makes this a very exciting time to delve into understanding what those models have learned and whether or not they learned about the 3D structure of the world in the process of learning to generate it. We hope that our findings will stimulate more interest in understanding the 3D awareness of visual models and that future work can provide more insight into how models represent the world and the impact of the learning objectives of such representations.

Acknowledgments: This work was done during an intership with the VisCAM team at Google Research. We thank Prafull Sharma, Shivam Duggal, Karan Desai, Junhwa Hur, and Charles Herrmann for many helpful discussions. We also thank Alyosha Efros, David Fouhey, Stella Yu, and Andrew Owens for their feedback.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2021. 1, 5, 7, 8, 15

[2] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *ICCV*, 2021. 3, 16

[3] Alexander C Berg, Tamara L Berg, and Jitendra Malik. Shape matching and object recognition using low distortion correspondences. In *CVPR*, pages 26–33. Citeseer, 2005. 7

[4] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, 2021. 3, 15

[5] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 14

[6] Anand Bhattad, Daniel McKee, Derek Hoiem, and DA Forsyth. Stylegan knows normal, depth, albedo, and more. *arXiv preprint arXiv:2306.00987*, 2023. 1, 4, 5, 8

[7] Thomas O. Binford. Visual perception by computer. In *Proceedings of the IEEE Conference on Systems and Control*, 1971. 2

[8] Rodney A Brooks. Symbolic reasoning among 3-d models and 2-d images. *Artificial intelligence*, 17(1-3):285–348, 1981. 2

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2, 3, 13, 21, 22, 23, 24, 25

[10] Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. In *ICML*, 2023. 1

[11] K Chatfield, K Simonyan, A Vedaldi, and A Zisserman. Return of the devil in the details: delving deep into convolutional nets. In *BMVC*, 2014. 8

[12] Yida Chen, Fernanda Viégas, and Martin Wattenberg. Beyond surface statistics: Scene representations in a latent diffusion model. *arXiv preprint arXiv:2306.05720*, 2023. 1, 4, 5, 8

[13] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 2, 5, 15

[14] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 13, 21, 22, 23, 24, 25

[15] 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 *Int. Conf. Learn. Represent.*, 2020. 2

[16] Xiaodan Du, Nicholas Kolkin, Greg Shakhnarovich, and Anand Bhattad. Generative models: What do they know? do they know things? let's find out! *arXiv*, 2023. 8

[17] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeruIPS*, 2014. 3, 15

[18] Mohamed El Banani and Justin Johnson. Bootstrap Your Own Correspondences. In *ICCV*, 2021. 5

[19] Mohamed El Banani, Luya Gao, and Justin Johnson. UnsupervisedR&R: Unsupervised Point Cloud Registration via Differentiable Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7129–7139, 2021. 5

[20] Mohamed El Banani, Karan Desai, and Justin Johnson. Learning Visual Representations via Language-Guided Sampling. In *CVPR*, 2023. 19

[21] 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(2):303–338, 2010. 15

[22] David Fouhey. *Factoring Scenes into 3D Structure and Style*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 2016. 3, 4

[23] David F. Fouhey, Wajahat Hussain, Abhinav Gupta, and Martial Hebert. Single image 3D without a single 3D image. In *ICCV*, 2015. 3, 16

[24] Micah Goldblum, Hossein Souri, Renkun Ni, Manli Shu, Viraj Prabhu, Gowthami Somepalli, Prithvijit Chattopadhyay, Mark Ibrahim, Adrien Bardes, Judy Hoffman, Rama Chellappa, Andrew Gordon Wilson, and Tom Goldstein. Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. In *NeurIPS Datasets and Benchmarks Track*, 2023. 1, 8, 13

[25] Walter Goodwin, Sagar Vaze, Ioannis Havoutis, and Ingmar Posner. Zero-shot category-level object pose estimation. In *ECCV*, 2022. 1, 8

[26] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *ICCV*, 2019. 1, 3, 8

[27] Kamal Gupta, Varun Jampani, Carlos Esteves, Abhinav Shrivastava, Ameesh Makadia, Noah Snavely, and Abhishek Kar. Asic: Aligning sparse in-the-wild image collections. *arXiv preprint arXiv:2303.16201*, 2023. 1

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015. 2

[29] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2, 3, 13, 21, 22, 23, 24, 25

[30] 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. 1, 8

[31] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave,

Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. 2, 14, 21, 22, 23, 24, 25

[32] Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, André Araujo, Ricardo Martin-Brualla, Kaushal Patel, Daniel Vlasic, Vittorio Ferrari, Ameesh Makadia, Ce Liu, Yuanzhen Li, and Howard Zhou. Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. In *NeurIPS Datasets and Benchmarks Track*, 2023. 2, 4, 14

[33] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. *arXiv preprint arXiv:2312.02145*, 2023. 8

[34] Gyeongnyeon Kim, Wooseok Jang, Gyuseong Lee, Susung Hong, Junyoung Seo, and Seungryong Kim. Dag: Depth-aware guidance with denoising diffusion probabilistic models. *arXiv preprint arXiv:2212.08861*, 2022. 8

[35] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 3

[36] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 3, 14, 21, 22, 23, 24, 25

[37] Jan J Koenderink and Andrea J Van Doorn. The singularities of the visual mapping. *Biological cybernetics*, 24(1):51–59, 1976. 8

[38] Jan J Koenderink and Andrea J Van Doorn. The internal representation of solid shape with respect to vision. *Biological cybernetics*, 32(4):211–216, 1979. 2, 8

[39] Jan J Koenderink and Andrea J Van Doorn. Surface shape and curvature scales. *Image and vision computing*, 10(8): 557–564, 1992. 2, 3

[40] Jan J Koenderink, Andrea J Van Doorn, and Astrid ML Kappers. Pictorial surface attitude and local depth comparisons. *Perception & Psychophysics*, 58(2):163–173, 1996. 4

[41] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019. 3, 8

[42] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. 3

[43] L'ubor Ladickỳ, Bernhard Zeisl, and Marc Pollefeys. Discriminatively trained dense surface normal estimation. In *ECCV*, 2014. 4, 14

[44] Martha Lewis, Qinan Yu, Jack Merullo, and Ellie Pavlick. Does clip bind concepts? probing compositionality in large image models. *arXiv preprint arXiv:2212.10537*, 2022. 5, 8, 17

[45] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2206–2217, 2023. 2

[46] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, 2021. 1

[47] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 15

[48] Zhuowan Li, Cihang Xie, Benjamin Van Durme, and Alan Yuille. Localization vs. semantics: Visual representations in unimodal and multimodal models. In *EACL*, 2024. 5, 8, 17

[49] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *ICLR*, 2020. 16

[50] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023. 1, 8

[51] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 14, 21, 22, 23, 24, 25

[52] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 3

[53] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 16

[54] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *arXiv*, 2023. 8

[55] David Marr and Tomaso Poggio. A computational theory of human stereo vision. *Royal Society of London*, 1979. 2

[56] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 1

[57] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv prepreint arXiv:1908.10543*, 2019. 2, 5, 15, 17

[58] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pretraining. In *Eur. Conf. Comput. Vis.*, 2022. 19

[59] J Farley Norman, James T Todd, Hideko F Norman, Anna Marie Clayton, and T Ryan McBride. Visual discrimination of local surface structure: Slant, tilt, and curvedness. *Vision research*, 46(6-7):1057–1069, 2006. 4

[60] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and

Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, 2023. 2, 3, 8, 13, 15, 18, 21, 22, 23, 24, 25

[61] Luigi Piccinelli, Christos Sakaridis, and Fisher Yu. idisc: Internal discretization for monocular depth estimation. In *CVPR*, 2023. 16

[62] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2022. 8

[63] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 8

[64] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 21, 22, 23, 24, 25

[65] 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 *Int. Conf. Machine Learning*, 2021. 1, 2, 3, 14

[66] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, Yuanzhen Li, and Varun Jampani. Dreambooth3d: Subject-driven text-to-3d generation. *ICCV*, 2023. 8

[67] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020. 3, 14

[68] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 3, 14, 15, 21, 22, 23, 24, 25

[69] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 3, 14, 21, 22, 23, 24, 25

[70] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1

[71] Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David Forsyth, and Anand Bhattad. Shadows don't lie and lines can't bend! generative models don't know projective geometry...for now. 2023. 8

[72] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 5, 15

[73] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 14

[74] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Int. Conf. Comput. Vis.*, 2017. 2, 8

[75] Roger N Shepard and Susan Chipman. Second-order isomorphism of internal representations: Shapes of states. *Cognitive psychology*, 1(1):1–17, 1970. 2

[76] Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971. 2

[77] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 8

[78] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2, 4, 13, 14

[79] Elizabeth Spelke, Sang Ah Lee, and Véronique Izard. Beyond core knowledge: Natural geometry. *Cognitive science*, 34(5):863–884, 2010. 2, 3

[80] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021. 14, 18

[81] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022. 5

[82] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 5

[83] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023. 1, 5, 7, 8, 14, 17

[84] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, 2019. 8

[85] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022. 8

[86] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit III: Revenge of the ViT. In *ECCV*, 2022. 2, 3, 13, 21, 22, 23, 24, 25

[87] Matthew Walmer, Saksham Suri, Kamal Gupta, and Abhinav Shrivastava. Teaching matters: Investigating the role of supervision in vision transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 1, 6, 8, 15

[88] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting

pretrained 2d diffusion models for 3d generation. In *CVPR*, 2023. 8

[89] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 13

[90] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, 2020. Association for Computational Linguistics. 13

[91] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *CVPR*, 2023. 13, 21, 22, 23, 24, 25

[92] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014. 15

[93] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views. In *CVPR*, 2023. 8

[94] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 3

[95] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. ODISE: Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. In *CVPR*, 2023. 1, 2, 14

[96] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *ICCV*, 2023. 15

[97] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *ICCV*, 2023. 3, 14, 21, 22, 23, 24, 25

[98] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. What does stable diffusion know about the 3d scene?, 2023. 8, 14, 15

[99] 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. In *NeurIPS*, 2023. 1, 5, 7, 8, 14, 15, 17

[100] Kaifeng Zhang, Yang Fu, Shubhankar Borse, Hong Cai, Fatih Porikli, and Xiaolong Wang. Self-supervised geometric correspondence for category-level 6d object pose estimation in the wild. In *ICLR*, 2023. 8

[101] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *arXiv preprint arXiv:2303.02153*, 2023. 2, 8, 14

[102] 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. 2, 3, 13, 21, 22, 23, 24, 25