

Probing the Mid-level Vision Capabilities of Self-Supervised Learning

Xuwei Chen¹ Markus Marks² Zezhou Cheng¹
¹University of Virginia ²California Institute of Technology

Abstract

Mid-level vision capabilities — such as generic object localization and 3D geometric understanding — are not only fundamental to human vision but are also crucial for many real-world applications of computer vision. These abilities emerge with minimal supervision during the early stages of human visual development. Despite their significance, current self-supervised learning (SSL) approaches are primarily designed and evaluated for high-level recognition tasks, leaving their mid-level vision capabilities largely unexamined.

In this study, we introduce a suite of benchmark protocols to systematically assess mid-level vision capabilities and present a comprehensive, controlled evaluation of 22 prominent SSL models across 8 mid-level vision tasks. Our experiments reveal a weak correlation between mid-level and high-level task performance. We also identify several SSL methods with highly imbalanced performance across mid-level and high-level capabilities, as well as some that excel in both. Additionally, we investigate key factors contributing to mid-level vision performance, such as pretraining objectives and network architectures. Our study provides a holistic and timely view of what SSL models have learned, complementing existing research that primarily focuses on high-level vision tasks. We hope our findings guide future SSL research to benchmark models not only on high-level vision tasks but on mid-level as well.

1. Introduction

Mid-level vision capabilities [41] (Fig. 1a) play a vital role in human visual development and daily life activities. By the age of one, children can re-organize retinal images into objects (*i.e.*, *perceptual grouping*), construct the visual world in three dimensions (*i.e.*, *3D geometric understanding*), track moving objects (*i.e.*, *cross-view correspondence*), grasp and bite these objects without knowing their semantics, and purposefully navigate through the 3D world. Importantly, such mid-level perception is developed without supervision [32].

On the other hand, self-supervised learning (SSL) in computer vision has made significant strides [5–8, 25, 42,

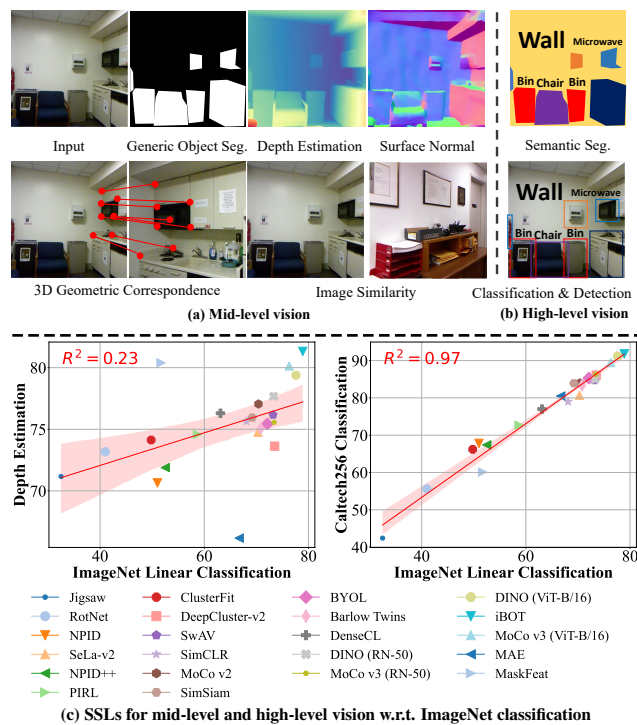


Figure 1. Mid-Level Vision vs. High-Level Vision Tasks. We provide a comprehensive evaluation of prominent self-supervised learning methods (SSLs) across a wide range of mid-level vision tasks (a), complementing the standard evaluation in high-level vision tasks (b). Although SSL performance in mid-level vision tasks (*e.g.*, depth estimation) is positively correlated with ImageNet linear probing, this correlation is much weaker than that observed among high-level vision tasks, as shown in (c) by the R^2 statistics.

45, 46, 66, 71]. The goal of SSL is to learn visual representations from unlabeled data so that these pretrained representations can be applied to a wide range of downstream tasks, reducing the need for human annotations and enhancing learning efficiency and performance. However, SSL is predominantly evaluated based on performance in high-level visual recognition tasks, such as image classification, object detection, and semantic segmentation (Fig. 1b). At the same time, the mid-level vision capabilities of SSL models are largely under-explored. To fill this gap, this pa-

per addresses the following questions:

- *Does strong high-level vision performance of SSLs indicate strong mid-level vision capabilities?*
- *What makes an SSL model effective for mid-level vision tasks?*

To answer these questions, we comprehensively evaluate SSLs in a wide range of mid-level vision tasks, including generic object segmentation, monocular depth and surface normal estimation, geometric correspondence, and mid-level image similarity (Fig. 1a and Tab. 2). We then relate the mid-level vision performance of these SSLs to their high-level vision performance. Our evaluations covers 22 seminar SSL approaches (Tab. 1) from diverse categories developed in the past decades (*e.g.* pretext tasks [25, 45], contrastive learning [6, 7, 9, 27, 66], self-distillation [5], and generative modeling [30, 62]). To make a controlled study and fair comparisons across different SSLs, our study primarily focuses on SSLs with publicly available checkpoints pretrained on ImageNet1K [12]. This allows us to gain insights into what factors may contribute to the mid-level vision capabilities in SSL (*e.g.*, training objectives, network architectures, *etc.*).

Our observations are as follows:

- The mid-level vision capabilities of SSLs are positively correlated with high-level vision. However, that correlation is generally weaker than that among high-level vision tasks (Fig. 1c and Fig. 2).
- Some SSLs have highly imbalanced mid-level and high-level vision capabilities. For instance, MAE [31] underperforms most SSLs in the mid-level vision tasks while having competitive high-level vision performance; MaskFeat [62] has better mid-level performance than most SSLs with similar high-level vision (Fig. 5).
- Network architectures, model capacity, and pretraining strategies impact mid-level vision capabilities. In general, ViT [14] outperforms ResNet [28]; Increasing the network capacity is beneficial (Tab. 3); The leading methods are generally based on generative modeling (Fig. 5).

Our study is built upon Probe3D [17], a recent work that evaluates the 3D awareness of visual models. In this work, we explore a broader range of mid-level vision tasks. Furthermore, our study primarily focuses on SSLs pretrained exclusively on ImageNet1K, while Probe3D probes visual foundation models that are trained from very diverse training configurations such as varying supervision signals (*e.g.* CLIP [50]) and dataset sizes. Our controlled settings allow us to gain insights into the factors contributing to model performance. Our exploration is also closely related to prior works that benchmark SSLs in various vision tasks [18, 26, 44, 55], such as out-of-domain classification [40] and fine-grained recognition [54, 57]. Unlike these works, our evaluation includes more recent SSLs, providing a timely study on this topic.

In summary, our study offers a comprehensive view of the representations learned by SSL models. We hope these findings inspire practical approaches that perform well across both mid-level and high-level vision tasks. We will release our benchmark suite to support future SSL benchmarking in the research community.

2. Related Work

Self-Supervised Learning (SSL). The goal of self-supervised learning (SSL) in computer vision is to pre-train image representations using unlabeled data, enabling these representations to be transferred effectively to various downstream tasks. Early SSL methods leveraged a range of pretext tasks, such as colorization [70], rotation prediction [25], jigsaw puzzle solving [45], and inpainting [47]. Recent advances have focused on contrastive learning [6, 7, 27, 29, 66, 69] and generative modeling [13, 31, 37, 48, 62]. Despite extensive research, SSL performance is primarily evaluated on high-level vision tasks (*e.g.*, ImageNet classification, object detection), overlooking the practical relevance of mid-level vision tasks. In this work, we focus on evaluating SSL methods on mid-level vision tasks.

Self-Supervised Pretraining for 3D Vision. Unlike most SSL methods focusing on 2D visual recognition tasks, a few recent works explore SSLs for 3D vision. For example, CroCo [63] adopts masked image modeling to learn 3D representations from image pairs showing the same scene from different viewpoints. Despite the improvement in downstream 3D vision tasks, CroCo compromises high-level recognition performance. As a followup, CroCo-v2 [64] extends the training data of CroCo to a large dataset of natural images and demonstrates improved performance in downstream optical flow and stereo vision tasks. However, learning representations for 3D vision tasks remains under-explored.

Benchmarking Self-Supervised Learning. Numerous studies underscore the growing importance of empirically evaluating general-purpose representation learning. For example, Goyal *et al.* [26] provide extensive benchmarking of SSLs across multiple downstream tasks, including classification, navigation, object detection, and surface normal estimation. Newell *et al.* [44] examine the effectiveness of SSLs under various configurations (*e.g.*, annotation quantities) in downstream tasks such as pose estimation and depth estimation. Ericsson *et al.* [18] evaluate 13 SSLs across 40 downstream vision tasks, demonstrating a weak correlation between SSL performance on ImageNet classification and tasks like few-shot learning, object detection, and dense prediction. More recently, Marks *et al.* [40] studied how classification-based SSL evaluation protocols correlate with and predict downstream performance on out-of-domain datasets. In line with these works, we comprehen-

sively evaluate existing SSLs across various downstream tasks. We focus on mid-level vision performance and how it relates to high-level vision.

Benchmarking Visual Foundation Models. More recently, several studies have evaluated large visual foundation models (VFMs) that are not exclusively trained by self-supervised learning methods (*e.g.*, CLIP [50]). For example, Probe3D [17] examines the 3D awareness of VFMs [5, 34], while Cambrian-1 [55] utilizes large language models (LLMs) and visual instruction tuning to assess visual representations across various 2D and 3D vision tasks. Since VFMs are typically trained with diverse approaches (*e.g.*, different supervision signals, datasets, and architectures), isolating the factors contributing to the performance of the representations they learn is challenging. In contrast, we conduct a controlled evaluation of SSL models that provide publicly available checkpoints pretrained on ImageNet1K. This approach enables fair comparisons across SSLs and offers deeper insights into their methodologies.

3. Background: Vision Task Taxonomy

There exist numerous computer vision tasks [68] which can be categorized from different perspectives. We briefly introduce two popular frameworks in this section:

- **Bottom-up Hierarchical Vision.** David Marr [41] proposes to break up the vision system into low-level, mid-level, and high-level stages. Low-level vision tasks aim to extract edges, contours, and blobs from raw pixels, while high-level vision concerns the semantics, functionality, and holistic 3D shape of objects. Mid-level vision, lying in-between the low-level and high-level vision, addresses the view-centric geometric properties of surfaces (*e.g.*, depth and surface normal), generic object segmentation and localization *without semantics*, cross-view image correspondences, and so on. This framework favors a feed-forward and bottom-up vision process.
- **Three R’s of Vision.** Malik *et al.* [39] categorize vision tasks into recognition, reconstruction, and reorganization. Recognition refers to the task of assigning semantic categories to images. Reconstruction concerns the estimation of 3D structures. Reorganization addresses the grouping and segmentation of images on spatial or perceptual similarity. Instead of a bottom-up process, 3R’s emphasizes the mutual benefits and relationship across these three visual capabilities.

Both frameworks highlight the importance of the non-semantic vision tasks — mid-level vision in Marr’s framework, reconstruction and reorganization in 3R’s. We adopt the term “mid-level” for the convenience of differentiating

from high-level recognition tasks. However, the underlying vision mechanism remains under debate.

4. Experimental Setup

In this section, we first introduce the SSLs considered in this work (Tab. 1) and our generic evaluation protocols for probing these SSLs. We then present the mid-level vision tasks in detail, including their definitions, datasets, and evaluation approaches (Sec. 4.1-4.3 and Tab. 2).

Our probing approaches are largely inspired by Probe3D [17], the recent work that evaluates the 3D awareness of visual foundation models (VFMs). The VFMs are usually trained under distinct settings (*e.g.* supervision signals, dataset sizes, network architectures and capacity), while we aim at probing self-supervised visual models that are trained with similar configurations. Therefore, our work provides a complementary study, specifically a more controlled setting, compared to Probe3D [17].

Self-Supervised Models. Tab. 1 presents an overview of the SSL models considered in our study. We primarily focus on SSLs that have publicly available checkpoints pretrained on ImageNet1K [12], with standard neural network architectures (*e.g.* ResNet50 [28] or Vision Transformers (ViT) [14]) as the image feature extractor. These SSLs cover a wide range of pretraining strategies, including clustering, contrastive learning, generative modeling, or a mixture of different training objectives. Targeting a comprehensive benchmark, we track the SSL literature back to classic methods that learn image representations from pre-text tasks such as Jigsaw [45] and RotNet [25]. Our selection of SSLs is inspired by Marks *et al.* [40], which provides a detailed study of the evaluation protocols of SSLs.

Generic Evaluation Protocols. Many mid-level vision tasks can be formulated as dense prediction (*e.g.*, depth estimation). Unlike image classification, dense prediction requires local pixel or patch representations as input, and the main design choice is to decide which network layers (*e.g.*, ResNet layers or ViT blocks) to extract the dense features from. Following Probe3D, we use the DPT decoder [52], a non-linear dense multi-scale network. For ResNet [28], we use the outputs of the last four ResNet blocks as the input to the DPT head, following the architecture of the feature pyramid network [36]; For ViT [14], following Probe3D, the network is split into four blocks evenly, and the features are extracted after each block — for ViT-B, we extract features after layers 3, 6, 9, and 12. Like Probe3D, we probe the frozen representations to evaluate the capabilities of learned representations instead of their transferability. We provide task-specific evaluation protocols in the following sections and hyperparameter settings in the supplementary material.

ID	SSL	Architecture	Category
1	Jigsaw [45]	ResNet50	Pre-text
2	RotNet [25]	ResNet50	Pre-text
3	NPID [66]	ResNet50	Contrastive
4	NPID++ [42]	ResNet50	Contrastive
5	PIRL [42]	ResNet50	Contrastive
6	ClusterFit [67]	ResNet50	Clustering
7	SwAV [4]	ResNet50	Clustering
8	SeLa-v2 [4]	ResNet50	Clustering
9	DeepCluster-v2 [4]	ResNet50	Clustering
10	SimCLR [6]	ResNet50	Contrastive
11	MoCo-v2 [8]	ResNet50	Contrastive
12	SimSiam [7]	ResNet50	Contrastive
13	DenseCL [58]	ResNet50	Contrastive
14	BYOL [27]	ResNet50	Contrastive
15	Barlow Twins [69]	ResNet50	Contrastive
16	MoCo-v3 [9]	ResNet50	Contrastive
17	DINO [5]	ResNet50	Self-distillation
18	MoCo-v3 [9]	ViT-B/16	Contrastive
19	DINO [5]	ViT-B/16	Self-distillation
20	MAE [31]	ViT-B/16	Generative
21	iBOT [71]	ViT-B/16	Mixture
22	MaskFeat [62]	ViT-B/16	Generative

Table 1. **Evaluated Self-Supervised Learning Methods (SSLs).** We provide a comprehensive evaluation of SSLs that provide publicly available checkpoints trained from the ImageNet1K dataset [12]. We categorize SSLs based on their training objectives and network backbones. *Mixture* denotes the combination of different SSLs (e.g., iBOT [71]). The SSLs are listed in order of their publication year.

4.1. Generic Object Segmentation

Generic object segmentation (or figure-ground segmentation) refers to the task of separating objects (figure) from the surrounding background *without any semantics*. This mid-level vision task is different from semantic segmentation (i.e., assigning each pixel to a *semantic category*) which is commonly used for evaluating the high-level visual capabilities of SSLs.

We use the VOC07 and VOC12 datasets [19, 20] as our benchmarks, following prior research on unsupervised object segmentation [56, 59, 61]. Given a pretrained feature extractor, we freeze the backbone and train a DPT decoder [51] with a simple binary cross-entropy loss (BCE). We use standard segmentation metrics (e.g., mean IoU) to measure the performance of SSLs in this task.

4.2. Monocular Surface Reconstruction

We evaluate SSLs in two tasks related to monocular 2.5D surface reconstruction: depth and surface normal estimation from Probe3D [17]. Even though these two tasks are closely related, they rely on different visual cues and cap-

ture distinct geometric properties.

We evaluate depth estimation and surface normal prediction at both scene and object levels. For scene-level performance, we employ the NYUv2 dataset [43], a widely used benchmark for indoor environments. Object-level depth is assessed using the NAVI dataset [33], which provides diverse object instances across scenes and orientations.

We present the task-specific configurations as follows.

Monocular Depth Estimation The task aims at predicting pixel-wise depth from monocular images. We adopt the formulation of binned depth estimation from AdaBins [21], instead of the classical regression approaches [15]. We estimate the scale-invariant depth for object-centric images (e.g., NAVI) while predict metric depths on scene-centric datasets (e.g., NYUv2). We report root-mean-squared error (RMSE) and recall at various threshold ratios.

Surface Normal Estimation. This task is to predict per-pixel surface direction. We train the DPT head with the uncertainty-aware angular loss [2]. As for depth estimation, we report RMSE and percentage recall at various angular thresholds.

4.3. Multiview Correspondence

In this section, we consider the tasks of estimating correspondences across multiple views that involve mid-level variations, such as changes in camera perspective and object motion. These mid-level variations differ from low-level variations in image properties (e.g., color and texture) and high-level variations in semantic attributes (e.g., object categories). Specifically, we probe SSLs in two tasks: geometric correspondence and mid-level image similarity. Following Probe3D [17], we extract image or pixel representations from pretrained SSLs model without additional training.

Geometric Correspondence This task aims to build pixel correspondences across two views that depict the same 3D scene. Our evaluation follows Probe3D [17]. Specifically, we estimate the pixel correspondences using the dense feature maps extracted from pretrained SSL models. We use the Paired ScanNet [10] to evaluate scene-centric images and the NAVI dataset [33] for object-centric images. We sample views with a maximum rotation of 120 degrees to ensure that a mutually visible surface exists without a large viewpoint change. We report the correspondence recall — the percentage of correspondence that falls within some defined distance.

Mid-Level Image Similarity The task aims at measuring the similarity of two images with mid-level variations (e.g., viewpoint). We evaluate the SSLs on the NIGHTS dataset collected by DreamSim [23]. Specifically, for each triplets $(x, \tilde{x}_0, \tilde{x}_1)$ from the NIGHTS dataset, we predict whether $\tilde{x}_0$ or $\tilde{x}_1$ is more similar to the reference image x . We denote a distance between two images as $D(x, \tilde{x}; f_\theta) =$

Hierarchy	Tasks	Type	Benchmarks	Losses	Metrics
High-level	Image Classification	2D	ImageNet1K [53]	CE	Accuracy
	Object Detection	2D	VOC07 [19], COCO [35]	CE+L2	AP
	Semantic Segmentation	2D	COCO [35]	CE	mIOU
Mid-level	Generic Object Segmentation	2D	VOC07 [19], VOC12 [20]	BCE	F1, mIoU, Acc
	Depth Prediction	3D	NAVI [33], NYU [43]	AdaBin	RMSE, ϵ_1 , ϵ_2 , ϵ_3
	Surface Normal Estimation	3D	NAVI [33], NYU [43]	Angular	RMSE, ϵ_1 , ϵ_2 , ϵ_3
	Geometric Correspondence	3D	NAVI [33], ScanNet [11]	Training-free	2D, 3D Recall
	Mid-level Image Similarity	2D	NIGHTS [24]	Training-free	Retrieval Score (F1-Score)

Table 2. **Probing Tasks.** In comparison with the standard SSL evaluation protocols that primarily focus on high-level vision tasks (**top**), we provide a comprehensive, systematic, and controlled evaluation of SSL in mid-level vision tasks (**bottom**). Our evaluations cover 2D and 3D vision with a wide range of benchmarks. The ϵ_1 , ϵ_2 and ϵ_3 in depth and surface normal estimation represents recall at different threshold ratios [16, 22]. Detailed descriptions of each metric can be found in the supplementary.

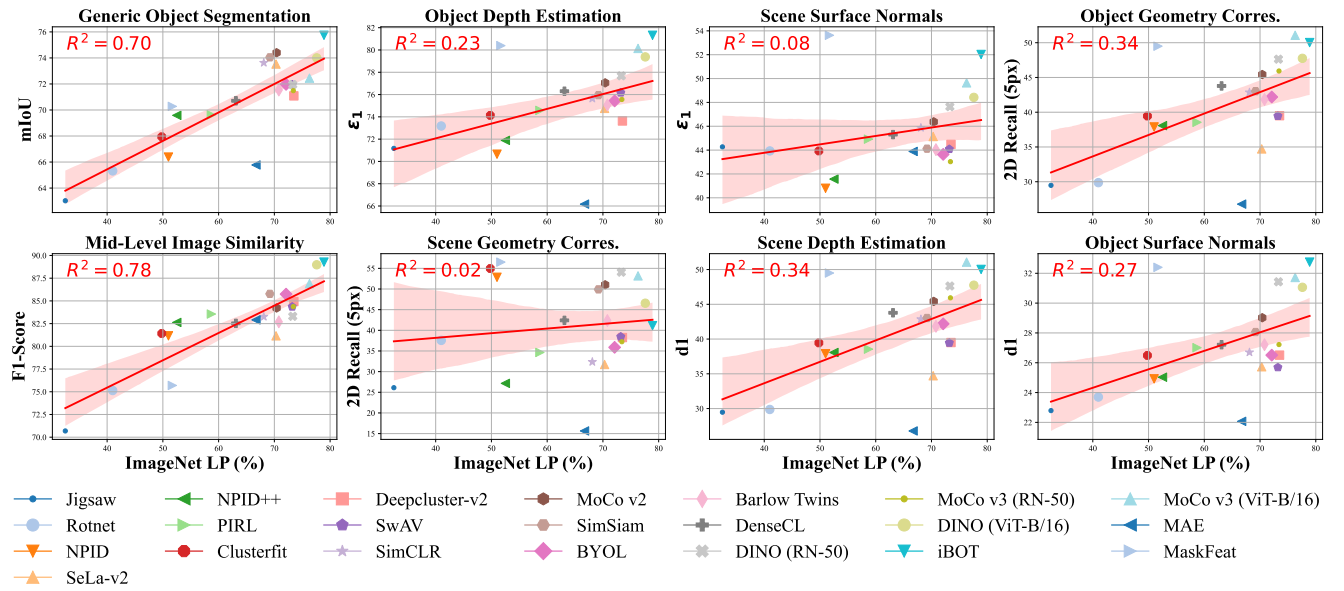


Figure 2. **Mid-Level Vision v.s. ImageNet Linear Probing.** We report the performance of self-supervised learning methods on mid-level vision tasks (y-axis) against their ImageNet 1K linear classification accuracy. Metrics are detailed in Tab. 2. Linear regression shows correlation with R^2 in each plot’s top-left corner, and the red shaded area marks the 95% confidence interval.

$1 - \cos(f_\theta(x), f_\theta(\tilde{x}))$ where f_θ is the image representations from pretrained SSLs. For ViT, we use the CLS tokens taken from the last layer; For ResNet, we use the global image representation in the output layer. The model’s vote is calculated as follows: if $d_1 < d_0$, then $\hat{y} = 1$; otherwise, if $d_0 < d_1$, then $\hat{y} = 0$, where $d_0 = D(x, \tilde{x}_0; f_\theta)$ and $d_1 = D(x, \tilde{x}_1; f_\theta)$. We use voting accuracy as our evaluation metric.

5. Analysis

5.1. Does strong high-level performance imply strong mid-level performance?

Yes, to some extent. As ImageNet (IN1k) probing results improve, we generally observe a corresponding enhance-

ment in mid-level vision tasks, as depicted in Fig. 2. This positive trend indicates that SSL models capable of high-level visual recognition tend to provide better representations for mid-level vision tasks as well, benefiting tasks that leverage spatial and structural information.

Some mid-level tasks are highly correlated with high-level. A per-task closer examination, however, reveals nuances in this relationship. For example, *generic object segmentation* shows the highest correlation with high-level tasks, achieving a coefficient of determination ($R^2 = 0.70$). This suggests that high-level features indeed capture spatial structures that enhance segmentation accuracy. Mid-level and high-level vision also correlate highly in the task

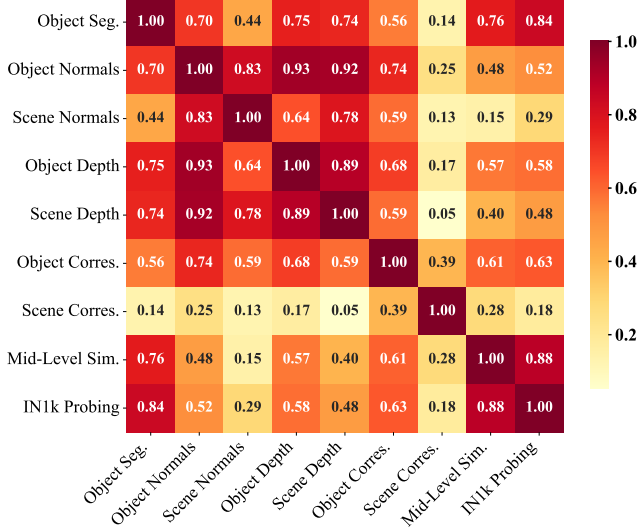


Figure 3. **Cross-task performance correlation.** We present Pearson coefficients across mid-level and high-level vision tasks. We use the same set of metrics we used in Fig. 2.

of *mid-level image similarity*. This indicates that the high-level image representations are invariant to mid-level variations to some extent (e.g., viewpoint variations).

3D geometric understanding is weakly correlated with high-level vision.

For tasks like *scene surface normal estimating*, we observe notably weaker correlations. These tasks often require intricate, fine-grained geometric and surface features that high-level SSL objectives alone do not sufficiently capture. This gap implies that mid-level vision depends on specialized cues not fully addressed by high-level feature extraction.

Correlation across mid-level vision tasks. Fig. 3 shows the Pearson correlations across tasks. We find a strong correlation between object segmentation and mid-level image similarity (coefficients above 0.76). Object segmentation further correlates well with depth estimation and object-level multi-view geometric correspondence, indicating that spatial understanding aids segmentation and underscores the importance of 3D awareness in image understanding. In contrast, scene geometry correspondence correlates weaker with other mid-level and high-level tasks. Mid-level image similarity aligns with object segmentation and correspondence but correlates less with 3D tasks.

Ranking in mid-level vision. In Fig. 5, we present a comparative ranking of SSL models based on their performance across mid-level vision tasks and ImageNet probing. Each model received a task-specific rank, and we computed an average score, with higher scores indicating stronger performance in both mid-level and high-level vision tasks. Our

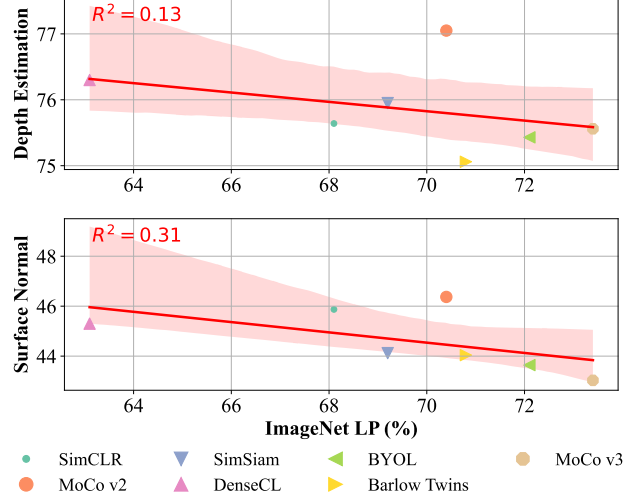


Figure 4. **Mid-Level Vision v.s. ImageNet Linear Probing with Contrastive Learning Methods.** We examine several contrastive learning methods [6–9, 27, 58, 69] and observe a negative correlation between ImageNet linear probing accuracy and task performance. Depth estimation results are shown at the top, with surface normal estimation results at the bottom. The y-axis represents prediction accuracy, where higher values indicate better performance.

analysis suggests an imbalance in model performance, with some models excelling in high-level tasks but showing comparatively modest capabilities in mid-level vision tasks. Notably, MaskFeat stands out as an exception, demonstrating a significant advantage in mid-level vision tasks over high-level performance.

Conversely, MAE ranks lower on mid-level tasks, demonstrating limited proficiency in capturing finer visual details. In the following sections, we will discuss the specifics of MAE and MaskFeat.

Notably, iBOT [71] and DINO [4] rank among the top performers, exhibiting strong capabilities in both mid-level and high-level vision tasks. In contrast, pretext pre-trained models like RotNet [25] and Jigsaw [45] fall behind other SSL methods.

5.2. What makes a strong mid-level vision model?

Training objective. We observe that the top-performing SSL models in mid-level vision tasks are typically generative-based, such as iBOT [71] and MaskFeat [62]. In contrast, recent contrastive self-supervised learning (SSL) models [6–9, 27, 58, 69] show an inverse relationship between mid-level vision performance and ImageNet linear probing accuracy, as illustrated in Fig. 4.

Although these methods have improved ImageNet linear probing accuracy by over 10%, they exhibit notable declines in mid-level vision tasks. We illustrate this trend with model performance on depth and surface normal estimation and

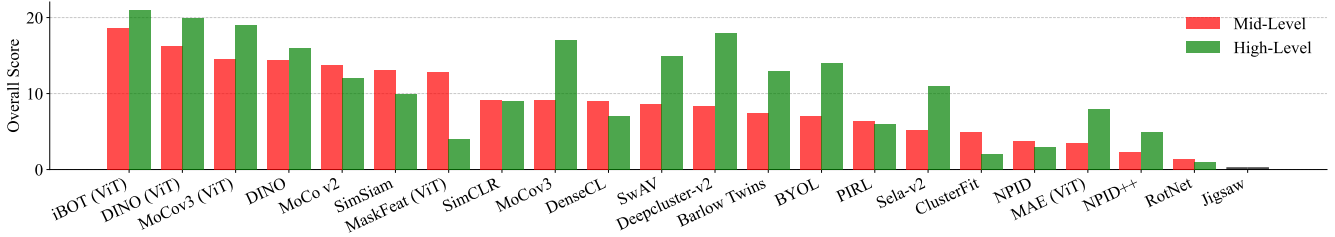


Figure 5. **Ranking SSLs based on Overall Score.** Comparative ranking of SSL models based on performance in mid-level vision tasks and ImageNet probing. Models are ranked per task, and average scores reflect overall performance across mid-level and high-level vision tasks, with higher scores indicating stronger results.

provide a broader analysis across six additional tasks in the supplementary materials. Notably, SSL models pre-trained on pretext tasks show the lowest performance on mid-level vision tasks.

Network architecture. We further compare ViT-B (ViT Base) and ViT-L (ViT Large) backbones with the same SSL method in Tab. 3. Our findings show that ViT-L models tend to perform better than ViT-B across mid-level vision tasks. In particular, we evaluate two popular SSL models, iBOT and MAE, on both ViT-B and ViT-L backbones. This pattern suggests that larger backbones like ViT-L may capture mid-level vision features better.

Network capability. We also compare ViT-B and ViT-L backbones using the same SSL method in Tab. 3. Our results indicate that ViT-Large models generally achieve higher scores across various mid-level vision tasks. Specifically, we examine two popular SSL models, iBOT and MAE, with ViT-B and ViT-L backbones. This trend suggests that larger models are better at capturing mid-level vision features.

5.3. Which SSL works the best per task?

In Tab. 4, we highlight the best-performing models for each task. iBOT [71] consistently leads in most mid-level vision tasks while MaskFeat performs even better on scene correspondence and scene surface normal. DINO works the best on object correspondence and achieves quite competitive results compared to other SSLs.

5.4. Analysis of specific SSL

MAE. MAE [30] stands out as a clear outlier, as illustrated in Fig. 2 and Fig. 5. MAE consistently performs poorly on mid-level vision tasks compared to other ViT-based SSLs such as iBOT [71] and MaskFeat [62].

MaskFeat. In Fig. 2 and Fig. 5, we observe that MaskFeat [62] demonstrates surprisingly strong mid-level vision capabilities despite showing relatively modest performance in high-level vision tasks. Since the MaskFeat checkpoint we used was trained exclusively on the ImageNet-1K dataset and followed a training process similar to MAE, we attribute its mid-level vision strength to its choice of target

features. Unlike models trained on raw pixels, tokens, or high-level features, MaskFeat is optimized on HOG (Histograms of Oriented Gradients) features, which are widely used for keypoint detection in methods such as SIFT [38]. Importantly, HOG does not rely on any external models, making it a robust feature for capturing mid-level vision attributes. This finding suggests that incorporating mid-level vision-relevant features, like HOG, into SSL training may enhance model performance on mid-level vision tasks.

SSLs based on pre-text tasks. RotNet and Jigsaw are the two lowest-performing SSLs on mid- and high-level vision. Given that they are the early design of SSLs, it re-validate our claim that ImageNet (IN1k) probing results improve, we generally observe a corresponding enhancement in mid-level vision task performance.

5.5. How do visual foundation models (VFMs) perform?

Even though our study aims to conduct a controlled benchmarking of SSLs in the mid-level vision tasks, we put VFMs into our context in this section to understand the comparison between the latest visual foundation models and the set of SSLs we analyzed. We observed that CLIP is much worse than most of the SSLs. DINO-v2 is the winner. We include a comparison in Fig. 6 focusing on depth estimation.

DINO-V2 [46]. Our findings, consistent with Probe3D and Cambrian-1 [55], show DINO-V2 as the best-performing model across both ImageNet linear probing and mid-level vision tasks, making it a standout choice for comprehensive vision capabilities.

CLIP [50]. CLIP’s overall performance is notably lower across mid-level tasks, as reported in Probe3D [17] and Cambrian-1 [55]. This suggests that, while generalizable for 2D understanding, CLIP may be less suited for tasks requiring 3D spatial understanding.

CroCo [65]. Using depth estimation as a representative task, we observe that CroCo performs competitively in mid-level vision tasks despite lagging in high-level vision tasks. Pretrained on synthetic multi-view image datasets, CroCo learns strong 3D-aware image representations. This distinction underscores its potential for 3D-specific applications,

SSLs	Backbone	Object Seg.	Object SNorm.	Scene SNorm.	Object Depth	Scene Depth	Object Corres.	Scene Corres.	Image Sim.
MoCo-v3 [9]	RN-50	71.48	27.22	43.03	45.93	75.56	1.7	15.23	84.37
MoCo-v3 [9]	ViT-B/16	72.45	31.72	49.64	51.07	80.14	3.12	25.03	86.90
DINO [5]	RN-50	71.95	31.43	47.64	47.63	77.68	3.16	26.63	83.31
DINO [5]	ViT-B/16	74.00	31.06	48.42	47.75	79.38	2.61	24.38	89.36

SSLs	Model Size	Object Seg.	Object Norm.	Scene Norm.	Object Depth	Scene Depth	Object Corres.	Scene Corres.	Image Sim.
iBOT	B/16	75.74	32.75	52.02	50.02	81.32	3.12	20.04	89.27
iBOT	L/16	75.32	35.65	54.53	52.56	85.14	3.16	14.79	82.37
MAE	B/16	65.77	22.67	43.89	26.78	66.17	2.74	6.64	82.91
MAE	L/16	66.58	26.60	45.60	26.81	70.22	2.93	10.05	84.26

Table 3. **Impact of Network Architecture and Model Size on Mid-Level Vision Performance.** The top table demonstrates that the Vision Transformer (ViT) [14] generally outperforms ResNet-50 (RN-50) [28] across most mid-level vision tasks when using MoCo-v3 or DINO as the self-supervised pretraining method. The bottom table shows that larger model sizes tend to enhance mid-level vision capabilities. Detailed numerical comparisons are provided in the supplementary material.

Model	Image Sim.	Object Corres.	Object Depth	Object SNorm.
DINO (RN-50)	83.31	3.16	47.63	31.43
MaskFeat [62]	75.70	2.08	<u>49.50</u>	<u>32.40</u>
MoCo v3	84.37	1.70	45.93	27.22
iBOT [71]	89.27	<u>3.12</u>	50.02	32.75

Model	Object Seg.	Scene Corres.	Scene Depth	Scene SNorm.
DINO (RN-50)	<u>71.95</u>	<u>26.63</u>	77.68	47.64
MaskFeat [62]	70.28	27.94	<u>80.39</u>	53.63
MoCo v3	71.48	15.23	75.56	43.03
iBOT [71]	75.74	20.04	81.32	<u>52.02</u>

Table 4. **Top SSL Methods in Mid-Level Vision Tasks.** iBOT outperformed all other SSL methods (listed in Tab. 1) in multiple mid-level vision tasks. Surprisingly, despite the low performance on ImageNet linear classification, MaskFeat achieved leading results in depth and surface normal estimation.

where spatial understanding is prioritized over high-level visual semantics because of its design choice.

SAM [34]. Although SAM excels at image segmentation, it exhibits moderate performance on mid-level vision tasks and notably weaker performance on image classification, in comparison with ImageNet-pretrained SSLs. This reveals the imbalanced performance of SAM among reorganization, reconstruction and recognition.

6. Discussion

In conclusion, this study introduces a suite of benchmark protocols to systematically evaluate mid-level vision capabilities, offering a comprehensive and controlled assessment of 22 SSL models across mid-level vision tasks. Our findings indicate a positive correlation between SSL performance in mid-level and high-level vision tasks, though this correlation is generally weaker than that observed within high-level tasks.

Some SSL models exhibit imbalanced performance across these task levels. For instance, MAE [31] underperforms in mid-level tasks despite competitive high-level performance, while MaskFeat [62] excels in mid-level tasks

¹ We evaluated CLIP with the ViT-B/16 checkpoint released by OpenAI.

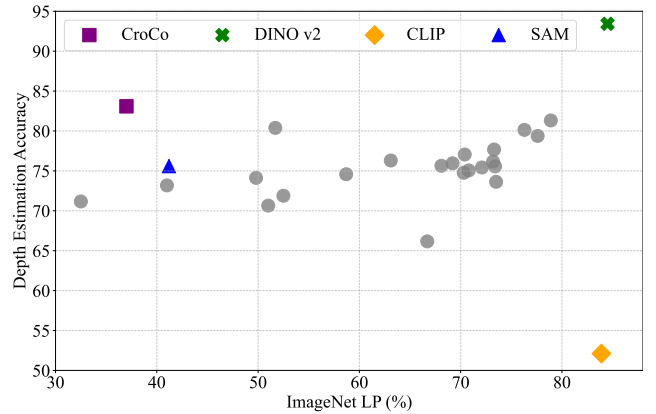


Figure 6. Comparison of ImageNet1K-Pretrained SSLs with models trained with larger datasets (e.g., DINO-v2 [46]) or different training supervisions (e.g., CLIP [50]¹) in the scene depth estimation task.

relative to its high-level results (Fig. 5). Mid-level vision capabilities are influenced by factors such as network architecture, dataset size, and pretraining strategy. Overall, ViT [14] outperforms ResNet [28], increasing model capacity proves beneficial (Tab. 3), and generative modeling approaches often lead in performance (Fig. 5). Additionally, we compare SSL models with VFMs, finding that DINOv2 [46] leads in both mid-level and high-level vision tasks and language objectives (e.g., CLIP) do not help on mid-level tasks.

Since SSL’s overall goal is to produce a very general feature representation that works well for a wide range of downstream tasks, we encourage the community to evaluate new SSL methods on both high- and mid-level tasks.

This work is limited by resources, which prevents us from scaling up our experiments to more tasks and models, and we did not include ending tasks such as navigation and robotic manipulation to emphasize mid-level vision’s role further. We provide details of our probing hyper-parameters and include visualizations for each task in the supplementary materials.

7. Acknowledgement

The authors gratefully acknowledge Research Computing at the University of Virginia for providing the computational resources and technical support that made the results in this work possible. [Research Computing at UVA](#).

References

- [1] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning, 2020. [2](#)
- [2] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation, 2021. [4](#)
- [3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features, 2019. [3](#)
- [4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2021. [4](#), [6](#), [2](#), [3](#), [8](#), [9](#), [10](#)
- [5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. [1](#), [2](#), [3](#), [4](#), [8](#), [9](#), [10](#)
- [6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *ICML*, 2020. [2](#), [4](#), [6](#), [1](#), [8](#), [9](#), [10](#)
- [7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. [2](#), [4](#), [8](#), [9](#), [10](#)
- [8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. [1](#), [4](#), [2](#), [8](#), [9](#), [10](#)
- [9] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021. [2](#), [4](#), [6](#), [8](#), [3](#), [9](#), [10](#)
- [10] 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. [4](#)
- [11] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes, 2017. [5](#)
- [12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. [2](#), [3](#), [4](#), [1](#)
- [13] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Maskclip: Masked self-distillation advances contrastive language-image pretraining, 2023. [2](#)
- [14] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. [2](#), [3](#), [8](#)
- [15] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Neural Information Processing Systems*, 2014. [4](#)
- [16] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network, 2014. [5](#), [4](#)
- [17] 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 *CVPR*, 2024. [2](#), [3](#), [4](#), [7](#)
- [18] Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How Well Do Self-Supervised Models Transfer? In *CVPR*, 2021. [2](#)
- [19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>, . [4](#), [5](#), [8](#)
- [20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>, . [4](#), [5](#), [8](#)
- [21] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021. [4](#)
- [22] David F. Fouhey, Wajahat Hussain, Abhinav Gupta, and Martial Hebert. Single image 3d without a single 3d image. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1053–1061, 2015. [5](#), [4](#)
- [23] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data, 2023. [4](#)
- [24] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *Advances in Neural Information Processing Systems*, 36, 2024. [5](#)
- [25] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. [1](#), [2](#), [3](#), [4](#), [6](#), [8](#), [9](#), [10](#)
- [26] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the IEEE/CVF International Conference on computer vision*, pages 6391–6400, 2019. [2](#)
- [27] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doherty, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. [2](#), [4](#), [6](#), [3](#), [8](#), [9](#), [10](#)

- [28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 3, 8
- [29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2
- [30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 2, 7, 3, 8, 9, 10
- [31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2, 4, 8, 9, 10
- [32] Donald D Hoffman. *Visual intelligence: How we create what we see*. WW Norton & Company, 2000. 1
- [33] Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, Andre Araujo, Ricardo Martin-Brualla, Kaushal Patel, Daniel Vlasic, Vittorio Ferrari, Ameesh Makadia, Ce Liu, Yanzhen Li, and Howard Zhou. Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. In *NeurIPS*, 2023. 4, 5
- [34] 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. In *ICCV*, 2023. 3, 8
- [35] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 5
- [36] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3
- [37] Yuan Liu, Songyang Zhang, Jiacheng Chen, Kai Chen, and Dahua Lin. Pixmim: Rethinking pixel reconstruction in masked image modeling, 2023. 2
- [38] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004. 7
- [39] Jitendra Malik, Pablo Arbeláez, Joao Carreira, Katerina Fragkiadaki, Ross Girshick, Georgia Gkioxari, Saurabh Gupta, Bharath Hariharan, Abhishek Kar, and Shubham Tulsiani. The three r’s of computer vision: Recognition, reconstruction and reorganization. *Pattern Recognition Letters*, 72:4–14, 2016. 3
- [40] Markus Marks, Manuel Knott, Neehar Kondapaneni, Elijah Cole, Thijs Defraeye, Fernando Perez-Cruz, and Pietro Perona. A closer look at benchmarking self-supervised pre-training with image classification. *arXiv preprint arXiv:2407.12210*, 2024. 2, 3
- [41] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010. 1, 3
- [42] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020. 1, 4, 2, 8, 9, 10
- [43] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *ECCV*, 2012. 4, 5
- [44] Alejandro Newell and Jia Deng. How useful is self-supervised pretraining for visual tasks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2
- [45] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 1, 2, 3, 4, 6, 8, 9, 10
- [46] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2023. 1, 7, 8
- [47] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 2
- [48] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers, 2022. 2, 4
- [49] Luigi Piccinelli, Christos Sakaridis, and Fisher Yu. idisc: Internal discretization for monocular depth estimation. In *CVPR*, 2023. 4
- [50] 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*. PMLR, 2021. 2, 3, 7, 8
- [51] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 4
- [52] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 3
- [53] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 5, 1, 2
- [54] Jong-Chyi Su, Zezhou Cheng, and Subhansu Maji. A Realistic Evaluation of Semi-supervised Learning for Fine-grained Classification. In *CVPR*, 2021. 2
- [55] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 2, 3, 7
- [56] Wouter Van Gansbeke, Simon Vandenhende, and Luc Van Gool. Discovering object masks with transformers for unsupervised semantic segmentation. *arXiv preprint arXiv:2206.06363*, 2022. 4

- [57] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisín Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893, 2021. [2](#)
- [58] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, 2021. [4](#), [6](#), [2](#), [8](#), [9](#), [10](#)
- [59] Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and Jose M Alvarez. Freesolo: Learning to segment objects without annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14176–14186, 2022. [4](#)
- [60] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3124–3134, 2023. [5](#)
- [61] Xudong Wang, Rohit Girdhar, Stella X. Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation, 2023. [4](#)
- [62] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022. [2](#), [4](#), [6](#), [7](#), [8](#), [3](#), [9](#), [10](#)
- [63] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *NeurIPS*, 35, 2022. [2](#)
- [64] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. CroCo v2: Improved Cross-view Completion Pre-training for Stereo Matching and Optical Flow. In *ICCV*, 2023. [2](#)
- [65] Weinzaepfel, Philippe and Leroy, Vincent and Lucas, Thomas and Brégier, Romain and Cabon, Yohann and Arora, Vaibhav and Antsfeld, Leonid and Chidlovskii, Boris and Csurka, Gabriela and Revaud Jérôme. CroCo: Self-Supervised Pre-training for 3D Vision Tasks by Cross-View Completion. In *NeurIPS*, 2022. [7](#)
- [66] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. [1](#), [2](#), [4](#), [8](#), [9](#), [10](#)
- [67] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations, 2019. [4](#), [1](#), [2](#), [8](#), [9](#), [10](#)
- [68] Amir R. Zamir, Alexander Sax, William B. Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. [3](#)
- [69] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021. [2](#), [4](#), [6](#), [3](#), [8](#), [9](#), [10](#)
- [70] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. [2](#)
- [71] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. [1](#), [4](#), [6](#), [7](#), [8](#), [2](#), [9](#), [10](#)