Understanding Geometric Representations in Self-Supervised Vision Transformers via Subspace Intervention

Weichen Zhou[1], Yawen Zou[1], Chunzhi Gu[2], Ran Dong[3], Haoran Xie[4], and Chao Zhang[1] *

[1] University of Toyama, Toyama, Japan
[2] University of Fukui, Fukui, Japan
[3] Chukyo University, Aichi, Japan
[4] Japan Advanced Institute of Science and Technology (JAIST), Ishikawa, Japan

Abstract. We introduce a controlled subspace intervention framework to investigate how self-supervised Vision Transformers (ViTs) encode dense geometric information. While linear probing is widely used to assess geometric representations, it treats features as a black box, failing to disentangle the underlying topology. To address this issue, we decompose the weights of converged linear probes to isolate the low-rank subspaces containing explicit geometric signals using Singular Value Decomposition (SVD). Our perspective yields three key insights: (1) Pre-training objectives determine how features are encoded. DINOv2 aligns spatial features for efficient linear extraction, while Masked Autoencoders (MAE) tend to disperse these signals, requiring a broader spatial context. (2) Explicit geometric representations are highly compressible, suggesting dense predictive heads could potentially be constrained to low-rank subspaces with minimal performance loss. (3) The layer-wise task affinity suggests that geometric precision peaks at intermediate layers before yielding to semantic abstraction in the final layers. By connecting internal encoding mechanics with downstream performance, these findings provide a basis for effective feature selection and lightweight decoder design. The source code is available at `https://github.com/Zhou-Weichen/Geosubprobe`.

Keywords: Representation Probing · Subspace Analysis · Task Affinity

1 Introduction

Modern dense prediction frameworks increasingly rely on self-supervised Vision Transformers (ViTs) as their foundational backbones [25, 39, 42, 43]. Beyond the expected gains from task-specific training, these models exhibit an inherent understanding of spatial geometric information directly within their pre-trained representation space [4, 30, 31]. While their empirical effectiveness is established, the encoding format of this emergent geometric information remains unclear.

* Corresponding author.

It is uncertain whether the network disperses geometric information across its entire high-dimensional feature space or aligns it into specific, linearly accessible coordinates. Drawing from insights that latent space topology is heavily influenced by pre-training objectives [33, 40, 41], we hypothesize that optimization constraints determine how geometric primitives are routed and formatted within latent subspaces. Understanding how distinct paradigms, such as self-distillation [31], masked image modeling [21], and hybrid approaches [47], shape these internal geometric representations is therefore critical for adapting these models to downstream tasks.

To characterize this inherent geometric understanding, recent studies reveal that self-supervised ViTs encode accurate single-view geometry despite lacking explicit 3D supervision [8, 13, 29, 44]. This geometric awareness is a functional necessity for resolving complex semantic ambiguities [45]. However, these models often struggle with multi-view consistency and complex spatial relationships [13, 29]. Such inconsistencies suggest that simply detecting the presence of geometric information is insufficient; it is crucial to understand how it is topologically encoded. Currently, representations are assessed primarily through the final prediction accuracy of black-box downstream probes [2, 8, 9, 13]. This paradigm conflates the absence of information with the inability to decode it [22,35]. When a linear probe performs poorly, it remains ambiguous whether the geometric information is absent, entangled within non-linear manifolds, or scattered across disjoint spatial patches. Resolving this diagnostic ambiguity requires directly examining the underlying feature encoding mechanics.

To address these uncertainties, we introduce a controlled subspace intervention framework. Since a linear probe extracts information by projecting features onto a learned weight matrix, its predictive capacity is bottlenecked by the continuous subspace spanned by these weights. Motivated by this property, we apply SVD to the converged weights of linear probes to explicitly isolate task-aligned geometric directions. This allows us to evaluate the encoding format where geometric signals recoverable from a low-dimensional subspace indicate highly aligned spatial features. In contrast, signals requiring high-dimensional non-linear aggregation across spatial tokens should be entangled.

Through this analytical framework, our investigation reveals distinct feature encoding formats driven by pre-training objectives. Self-distillation (*e.g.*, DINOv2) strongly aligns explicit geometry into highly compressible, low-rank coordinate systems. In contrast, generative masked reconstruction model (*e.g.*, MAE) disperses geometric signals across broader dimensions to satisfy pixel-level reconstruction constraints. Despite these differences, we identify a high compressibility of explicit geometric representations across all evaluated paradigms. Furthermore, layer-wise evaluation uncovers a task affinity among surface normal estimation, depth estimation, and semantic segmentation in the deeper representation space.

Our main contributions are summarized as follows:

– We introduce a controlled subspace intervention framework. By applying SVD to converged linear probe weights, we isolate task-aligned directions

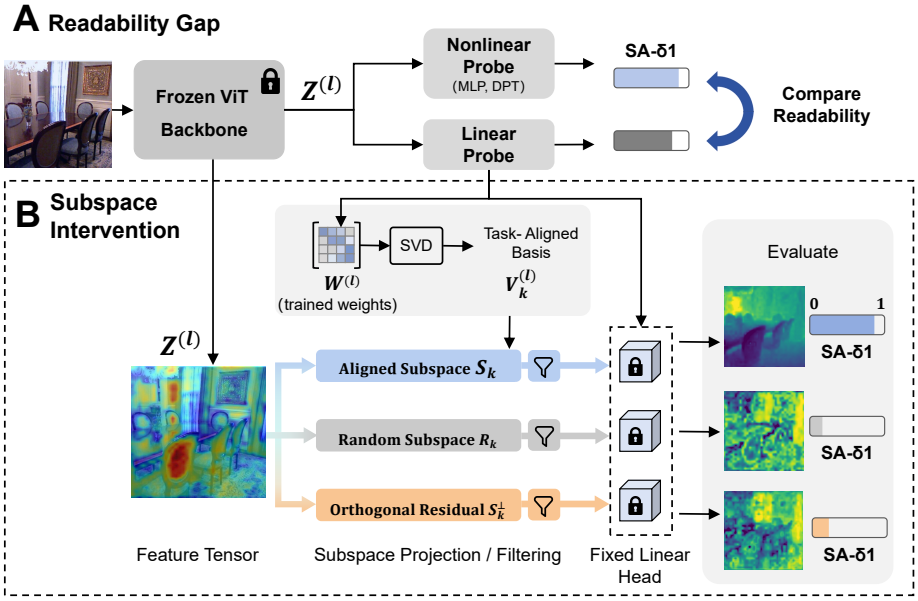


Fig. 1: Overview of the controlled subspace intervention analysis framework. (A) We evaluate frozen backbone features $\mathbf{Z}$ using three-tier probes (Linear, MLP, and DPT) to decouple local non-linear entanglement from global spatial fragmentation, thereby obtaining the readability gap of geometric features. (B) Without additional training, we perform SVD on the converged linear weights $\mathbf{W}$ to extract a task-aligned basis $\mathbf{V}_k$. The feature tensor is then projected onto the aligned subspace ($\mathcal{S}_k$), a random subspace ($\mathcal{R}_k$), and the orthogonal residual ($\mathcal{S}_k^{\perp}$). The projected features are evaluated through the fixed linear head to isolate the geometric signal.

to quantify the linear compressibility of geometric representations, moving beyond black-box output metrics.
- Through our framework, we characterize differences in the accessibility of geometric information across the evaluated self-supervised paradigms. DINOv2 exhibits more linearly accessible geometry than MAE, while much of the linearly decodable geometric information is concentrated in low-rank, task-aligned subspaces.
- We reveal that geometric features peak in DINOv2's middle layers, then weaken as semantic performance improves.

2 The Diagnostic Framework

Traditional evaluation paradigms assess representations through the black-box probing of downstream outputs, offering limited insights into the underlying feature structure. We introduce a post-hoc analytical framework that transitions from only evaluating task accuracy to inspecting the internal manifold topology.

As illustrated in Fig. 1, our approach consists of two primary stages: establishing a readability gap (Panel A) and performing subspace interventions (Panel B).

2.1 Readability Gap

Our first objective is to determine whether geometric information is genuinely absent or locked within complex non-linear structures. Rather than using a single decoder, we design a three-tier probing hierarchy. By deliberately restricting the capacity of these probe heads, we decouple local non-linear folding from the necessity of global spatial context. The resulting performance discrepancies across these tiers define the readability gap, providing a baseline of topological entanglement before we delve into its subspace.

***Tier1*: Linear Probing (Explicit Geometry).** Given the patch-level representation $\mathbf{Z}^{(l)} \in \mathbb{R}^{N \times D}$ from layer l of a frozen backbone, where N denotes the number of spatial tokens and D represents the feature dimension per token, we apply a linear head parameterized by $\mathbf{W}^{(l)} \in \mathbb{R}^{C \times D}$, where C is the target prediction space dimensionality (*e.g.*, $C = 256$ depth bins [10]). The prediction is computed as $\hat{\mathbf{Y}} = \mathbf{Z}^{(l)}\mathbf{W}^{(l)\top}$. This naturally extends to a concatenated multi-scale representation mapped by $\mathbf{W}_{\text{global}} \in \mathbb{R}^{C \times 4D}$. We treat this linear probe as a strict measure of accessibility, testing whether geometric cues are explicitly rectified into a linearly readable format without requiring cross-patch communication.

***Tier2*: MLP and *Tier3*: DPT Nonlinear Controls (Entanglement vs. Fragmentation).** When a linear probe fails, it remains unclear whether the geometry is entirely absent, non-linearly folded within individual patches, or scattered across the spatial manifold. To isolate this, we introduce a token-wise Multi-Layer Perceptron (MLP) and a DPT-style decoder [37]. The MLP operates point-wise on individual spatial tokens but incorporates non-linear activations, so the performance gap between the Linear and MLP probes quantifies the degree of local non-linear entanglement. The performance gap between the DPT and MLP probes isolates the degree of spatial dispersion, revealing whether the backbone distributes geometric primitives across multiple spatial tokens that necessitate a global receptive field for inference. Together, this probing hierarchy establishes a comprehensive readability description.

2.2 Subspace Intervention

While the readability gap quantifies the degree of entanglement, it does not reveal the directional distribution of information. To peer inside the linear probe's mechanism, we introduce a subspace intervention post-hoc analysis framework.

SVD on Probe Weights. We posit that the converged linear weight matrix $\mathbf{W}^{(l)}$ encodes the task-aligned geometric directions discovered by the probe. Because the mapping projects to a low-dimensional target space ($C \ll D$), the rank of the learned linear mapping is upper-bounded by C. We perform SVD directly on the probe weights without requiring additional training:

$$\mathbf{W}^{(l)} = \mathbf{U}^{(l)} \boldsymbol{\Sigma}^{(l)} \mathbf{V}^{(l)\top} \tag{1}$$

where $\mathbf{U}^{(l)} \in \mathbb{R}^{C \times C}$ contains the left singular vectors, $\boldsymbol{\Sigma}^{(l)} \in \mathbb{R}^{C \times C}$ is a diagonal matrix containing the singular values, and $\mathbf{V}^{(l)} \in \mathbb{R}^{D \times C}$ contains the right singular vectors. The columns of $\mathbf{V}^{(l)}$ form an orthonormal basis of the task-relevant representation space. By extension, the same decomposition is applied to the global concatenated weight matrix $\mathbf{W}_{\text{global}}$, yielding a corresponding low-rank basis spanning the multi-scale feature dimensions.

Aligned Subspace Identification. We define aligned subspace of rank k ($k \leq C$) as the span of the top-k principal directions:

$$\mathcal{S}_k^{(l)} = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}. \tag{2}$$

To assess the geometric density of this subspace, we constrain the backbone representation via orthogonal projection:

$$\tilde{\mathbf{Z}}_k^{(l)} = \mathbf{Z}^{(l)} \mathbf{V}_k^{(l)} \mathbf{V}_k^{(l)\top}, \tag{3}$$

Crucially, the projected features are evaluated directly using the original, frozen linear head. Because both the backbone and the probe remain locked, any performance variation is exclusively attributable to the geometric capacity of the selected k-dimensional subspace. To ensure that the extracted basis reflects the intrinsic manifold rather than optimization artifacts, we evaluated the similarity of the extracted subspaces across multiple random probe initializations (see Section 4.5).

2.3 Control Subspaces: Isolating the Signal

To ensure that the explicit geometric signal is directionally concentrated along the principal components, rather than an artifact of retaining k degrees of freedom, we introduce two control subspaces for strict ablation under the frozen probe:

First, the random subspace $\mathcal{R}_k$ projects features onto a span of k randomly sampled orthonormal directions in $\mathbb{R}^D$ or $\mathbb{R}^{4D}$. This disrupts the task-aligned structure while preserving the exact dimensionality, thereby serving as a null baseline. Second, to isolate the orthogonal complement containing the rejected tail dimensions, we compute the residual representation $\mathbf{Z}_{\text{res}}^{(l)}$ by subtracting the task-aligned projection from the original features:

$$\mathbf{Z}_{\text{res}}^{(l)} = \mathbf{Z}^{(l)} - \mathbf{Z}^{(l)} \mathbf{V}_k^{(l)} \mathbf{V}_k^{(l)\top}. \tag{4}$$

Passing these residual features through the frozen head allows us to validate the sufficiency of the low-rank subspace and assess the information loss in its orthogonal complement.

Evaluating these residual representations through the same frozen linear probe serves as a diagnostic for whether the task-aligned geometric signal is concentrated in the top-k subspace. A sharp decrease in performance on the residual space indicates that the explicitly decodable geometric features exploited by the probe are largely confined to the top-k components, rather than being broadly dispersed across the remaining dimensions. This experimental design determines whether the network compresses geometric signals into a compact manifold or relies on a more dispersed distribution.

3 Experimental Design and Objectives

3.1 Datasets and Implementations

We evaluate three representative self-supervised ViT paradigms: self-distillation (DINOv2 [31]), masked image modeling (MAE [21]), and a hybrid approach (iBOT [47]). For our primary analysis, we adopt the ViT-Large variant for all architectures and maintain frozen backbone weights.

Geometric representations are primarily assessed via monocular depth estimation on the standard NYU Depth V2 dataset [38]. Performance is measured by scale-aware (SA) threshold accuracy ($\delta < 1.25$), RMSE, and Absolute Relative error (AbsRel) [15]. To investigate layer-wise task affinities within a controlled domain, we perform parallel probing for 40-class semantic segmentation [11] and surface normal estimation, adopting the mean Intersection over Union (mIoU) and angular accuracy [6]($d_1 < 11.25°$) as the respective evaluation metrics.

Since one of the main contributions of this work is the analytical framework itself, which utilizes converged probe weights to inversely explore the internal feature topology, we focus on depth estimation to construct a cohesive and deep narrative. However, to examine the consistency of findings across tasks, domains, and probe configurations, we provide extensive parallel evaluations in the Supplementary Material. These include full surface normal trajectories, extended cross-domain benchmarks (*e.g.*, depth on KITTI [18, 19], and depth/normal estimation on NAVI [24]), and validation on ViT-Base variants.

3.2 The Three-tier Probing Hierarchy

As formalized in Sec. 2, we implement a three-tiered probing hierarchy to systematically distinguish the presence of geometric information from its structural accessibility. The baseline **Linear Probe** employs a 1×1 convolution to map frozen patch tokens directly to the target space, quantifying explicit, isolated geometric cues. To isolate local non-linear entanglement, the intermediate **MLP Probe** introduces point-wise non-linear activations (*e.g.*, a hidden layer with GELU) while strictly maintaining the exact same 1×1 token-wise receptive

Table 1: Linear Readability Gap Analysis. Comparing the geometric performance (SA-δ_1) across our three-tier probing. The decomposition of gaps isolates point-wise non-linear folding (Local Entanglement) from reliance on global context (Spatial Fragmentation), quantifying the degree of spatial feature alignment versus dispersion.

	Probe Architecture (SA-δ_1 ↑)			**Accessibility Gaps**	
Backbone	**Linear**	**1×1 MLP**	**DPT**	**Local Entang.**	**Spatial Frag.**
DINOv2-L	0.9157	0.9325	0.9483	+0.0168	+0.0158
iBOT-L	0.8198	0.8376	0.8524	+0.0178	+0.0148
MAE-L	0.6033	0.6390	0.7022	+0.0357	+0.0632

field. Finally, the **DPT Decoder** [37] establishes the absolute latent geometric upper bound by incorporating multi-scale feature aggregation and global spatial receptive fields.

3.3 Analytical Roadmap

Our experimental analysis follows a deductive trajectory, including each stage and its corresponding objective.

(1) Sec. 4.1: We assess the readability gap by applying different decoders (Linear, MLP, and DPT) to concatenated multi-layer features to quantify feature alignment versus dispersion.

(2) Sec. 4.2: Building upon the decoder baselines in (1), we apply SVD to the global probe weights to evaluate the compressibility of geometric representations and quantify their cross-layer energy routing.

(3) Sec. 4.3: Motivated by energy imbalances, we investigate rank sensitivity at individual network depths through single-layer subspace interventions.

(4) Sec. 4.4: We contrast geometric and semantic probing to examine whether terminal performance shifts originate from capacity loss or a transition in task-specific affinities.

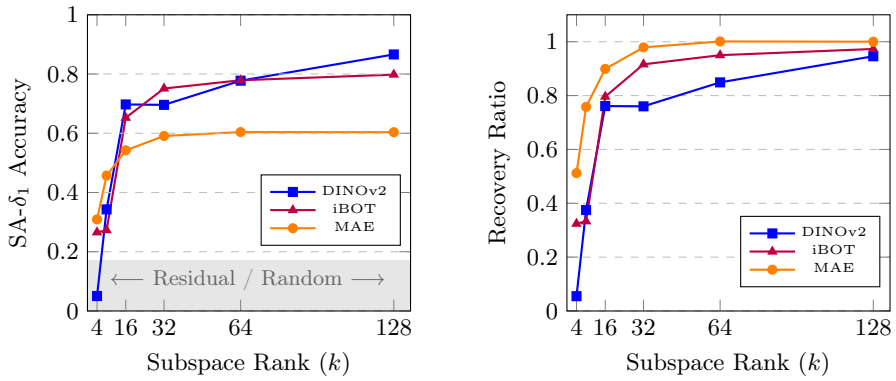
(5) Sec. 4.5: We verify the stability of the extracted subspaces across random probe initializations.

Full implementation details, including hyperparameters, optimizer, and exact architectural specifications, are provided in the Supplementary Material.

4 Results and Analysis

4.1 Feature Encoding : Alignment vs. Dispersion

As shown in Tab. 1, evaluating the three pre-training paradigms via our tiered probing reveals distinct feature encoding formats, distinguishing geometric representation presence from structural accessibility for downstream decoders.



(a) (b)

Fig. 2: (a) Absolute Recovery: Solid colored lines depict the performance of the task-aligned subspace ($\mathbf{V}_k\mathbf{V}_k^\top$). The gray shaded area denotes the noise floor (SA-$\delta_1 < 0.18$), accounting for both the residual subspace and the random orthogonal baselines under the frozen probe. This gap indicates substantial representational redundancy, suggesting that explicitly decodable geometric information can be compressed into a low-rank subspace without significant loss. **(b) Recovery Efficiency:** Normalized against each model's full-rank linear baseline, MAE (Orange) exhibits the fastest relative saturation, recovering $> 98\%$ of its linear potential by $k = 32$. iBOT (Purple) tracks closely with MAE, while DINOv2 (Blue) shows slower convergence. This indicates that DINOv2 linearly encodes richer, fine-grained geometric details that require slightly more dimensions to fully resolve.

DINOv2 exhibits strong spatial feature alignment. A linear probe alone recovers the vast majority of geometric signals (SA-δ_1 of 0.9157). Point-wise MLP probing improves this slightly to 0.9325, while the global DPT decoder yields 0.9483. This indicates that self-distillation inherently organizes geometric information into a linearly accessible format, minimizing the need for complex downstream decoders. In contrast, MAE demonstrates significant spatial dispersion. Linear performance drops to 0.6033, and MLP probing only raises accuracy to 0.6390. However, the DPT decoder's global receptive field substantially boosts performance to 0.7022. This jump (+0.0989 over Linear) indicates that masked reconstruction distributes geometric primitives across patches, necessitating global spatial aggregation for effective extraction. The hybrid iBOT model represents an intermediate state. Its minimal performance gains from the 0.8198 linear baseline to the 0.8524 DPT indicate low non-linear entanglement and fragmentation, suggesting that hybrid objectives balance linear accessibility with dense localized representations.

4.2 Global Compressibility and Energy Allocation

We apply subspace intervention on fused multi-layer representations to evaluate absolute saturation, relative structural efficiency, and energy allocation.

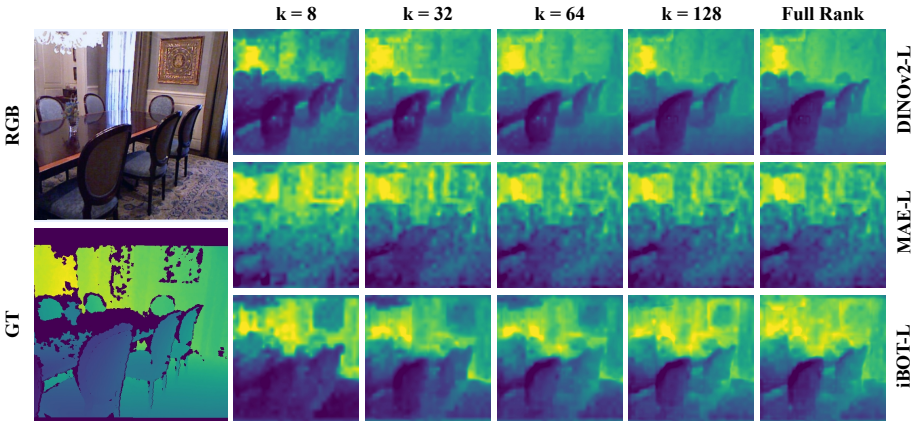


Fig. 3: Qualitative Visualization of Subspace Interventions. We show depth predictions from the global linear probe across selected ranks (k). The patchy artifacts result from applying linear projections directly to coarse patch tokens, a deliberate choice intended to expose the raw feature structure. DINOv2 (top) recovers coherent scene layouts at an extremely low rank ($k = 8$), demonstrating high linear compressibility. In contrast, MAE (middle) requires higher dimensional capacity ($k \geq 64$) to resolve basic object boundaries, as geometric signals are distributed among localized high-frequency details at low ranks. iBOT (bottom) represents an intermediate, hybrid state.

High Compressibility of Geometric Representations. We truncate the task-aligned basis to isolate explicitly encoded geometric signals. Fig. 2a demonstrates that performance on both random ($\mathcal{R}_k$) and residual ($\mathcal{S}_k^{\perp}$) control subspaces collapses under the frozen linear probe. This gap indicates significant representational redundancy, suggesting that explicitly decodable geometric information can be projected into a low-rank subspace with minimal loss. This performance collapse confirms that the predictive mechanism of the original probe is strictly bottlenecked within the top-k dimensions.

Qualitative visualizations (Fig. 3) and recovery efficiency curves (Fig. 2b) reveal distinct saturation behaviors across paradigms. MAE exhibits the fastest relative saturation, recovering over 98% of its linear potential at $k = 32$ despite its lower absolute geometric capacity. In contrast, DINOv2 requires a higher intrinsic dimensionality ($k \geq 64$) to converge. These results corroborate our accessibility findings. Specifically, MAE provides only coarse geometry in a linearly accessible form and requires few singular vectors for reconstruction, given that fine-grained details remain non-linearly entangled. In contrast, DINOv2 linearly aligns richer spatial details, which necessitates a broader basis to resolve precise object boundaries.

Cross-Layer Energy Allocation. While global subspace analysis quantifies the capacity required for geometry, it does not identify which layers contribute

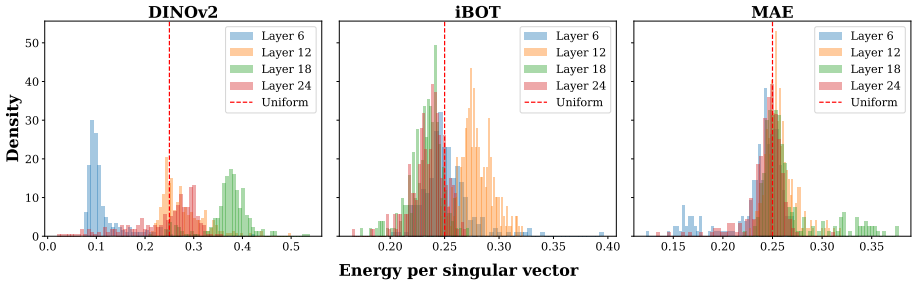


Fig. 4: Energy Distribution Across Singular Directions. Spectral-weighted energy density per singular vector. The dashed line represents the uniform baseline (0.25). DINOv2 shows multimodal peaks, indicating that geometric components are highly localized within intermediate layers. In contrast, iBOT and MAE exhibit overlapping distributions near the uniform baseline, indicating that their geometric signals are diffusely distributed across the hierarchy.

most to this signal. To determine the contribution of each layer, we analyze the energy distribution of the converged global linear probe via its singular spectrum.

The probe maps concatenated features using weights $\mathbf{W}_{\text{global}} \in \mathbb{R}^{C\times 4D}$. Each right singular vector $\mathbf{v}_m \in \mathbb{R}^{4D}$ resulting from the subspace decomposition encodes a task-aligned direction. We partition these vectors into four layer-specific blocks $\mathbf{v}_m = [\mathbf{v}_m^{(l_6)}, \mathbf{v}_m^{(l_{12})}, \mathbf{v}_m^{(l_{18})}, \mathbf{v}_m^{(l_{24})}]^\top$. To quantify the specific contribution of each depth, we define the spectral-weighted energy for layer i:

$$E_i = \frac{\sum_{m=1}^{C} \sigma_m^2 \|\mathbf{v}_m^{(l_i)}\|_2^2}{\sum_{j=1}^{4}\sum_{m=1}^{C} \sigma_m^2 \|\mathbf{v}_m^{(l_j)}\|_2^2} \tag{5}$$

where singular values σ_m weight each direction by its explanatory variance.

Spectral energy profiles (Tab. 2 and Fig. 4) reveal distinct routing strategies across depths. DINOv2 exhibits highly localized energy allocation, concentrating over 72% of geometric energy within intermediate layers (l_{12}, l_{18}) before a sharp decline to approximately 10% at l_{24}. This distribution suggests an empirical layer-wise affinity, where explicit geometric information is dominant in intermediate layers but declines in deeper ones. However, masked reconstruction and hybrid paradigms distribute this explicit signal. Both iBOT and MAE exhibit highly similar energy distributions that cluster near the uniform 0.25 baseline, requiring the linear probe to aggregate features across all depths simultaneously. This indicates that masked reconstruction distributes features across layers rather than concentrating them.

Table 2: Relative geometric contribution E_i for four representative depths (l_6 to l_{24}).

	Contribution (%)			
Model	l_6	l_{12}	l_{18}	l_{24}
DINOv2	17.2	35.8	**36.7**	10.3
iBOT	**26.9**	28.4	22.4	22.3
MAE	19.5	28.4	**32.7**	19.4

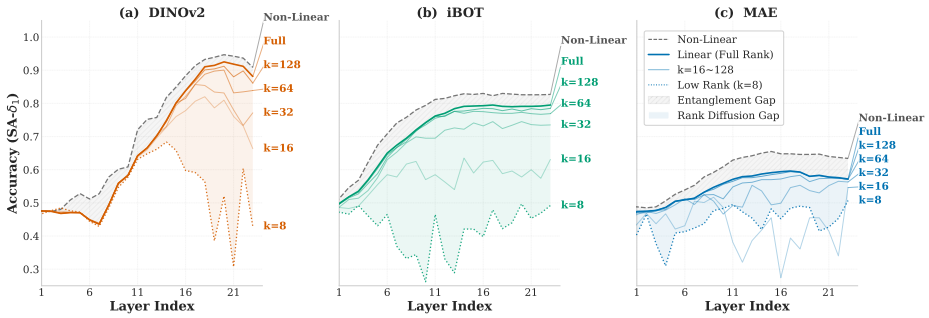


Fig. 5: Layer-wise Subspace Analysis. Performance evolution of geometric representations across different layers and subspace ranks. The Entanglement Gap highlights the disparity between linear and nonlinear decoding, indicating the degree of local nonlinear folding. The rank dispersion gap highlights the performance drop under extreme low-rank constraints ($k = 8$). Notice the distinct rank sensitivity transition in DINOv2 (a) at intermediate layers, contrasting with the continuous rank sensitivity in MAE (c) and iBOT (b).

This global aggregation inherently blends distinct layer-wise states. For example, the multi-scale probe compensates for the final-layer drop in DINOv2 by relying on intermediate features, masking the abrupt loss of geometric capacity. To trace how spatial understanding evolves and identify where it degrades, we therefore transition to a strictly single-layer diagnosis.

4.3 Layer-Wise Evolution and Rank Sensitivity

To trace the layer-wise evolution of geometric representations, we apply subspace interventions strictly to single-layer features. Fig. 5 tracks the performance trajectory from an extreme low-rank state ($k = 8$) to the full linear capacity, benchmarked against a non-linear upper bound. We establish this non-linear baseline using a lightweight decoder incorporating spatial convolutions and residual connections on isolated single-layer features. Full architectural details are provided in the Supplementary Material.

As shown in Fig. 5, DINOv2 exhibits significant geometric compactness in its early to middle layers. Up to layer 11, the $k = 8$ subspace performs nearly as well as the full-rank linear probe, indicating that self-distillation effectively aligns spatial primitives into a highly compressed, linearly accessible format. A distinct transition occurs near layer 12, where the $k = 8$ trajectory diverges from the full-rank baseline. This terminal decline in explicitly decodable geometry aligns with the localized intermediate energy routing observed in Sec. 4.2, implying a structural shift in the representational focus of the network. This transition motivates the subsequent analysis of layer-wise task affinities. In contrast, masked reconstruction distributes features across more dimensions. MAE shows immediate and persistent rank sensitivity, likely because pixel-level objectives preserve high-frequency textures that disperse geometric signals. Although

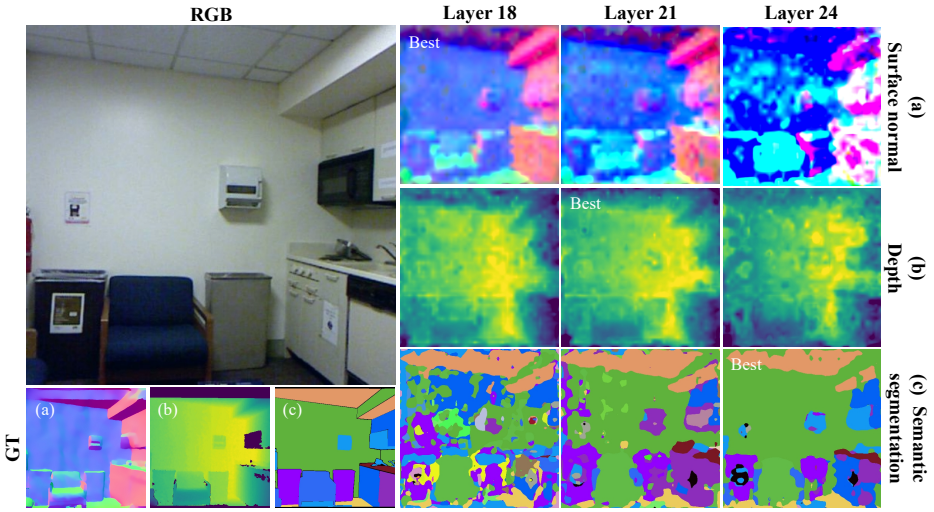


Fig. 6: Qualitative Task Affinity in DINOv2. Visual predictions decoded from frozen features at layers 18, 21, and 24. Performance for (a)surface normals, (b)depth, and (c)semantic segmentation peaks at these layers, respectively. This variation demonstrates distinct layer-wise affinities determined by the downstream task objectives.

iBOT exhibits some early-layer alignment, it does not achieve the low-rank compactness observed in DINOv2. Both masked reconstruction and hybrid models maintain a substantial entanglement gap across most depths, suggesting that their geometric information remains reliant on local non-linear structures.

Despite these divergent evolutionary paths, the $k = 64$ and $k = 128$ subspace trajectories closely track the full-rank linear baselines across nearly all depths in every paradigm. This consistency suggests a pervasive high compressibility, where explicitly decodable geometric representations can be recovered from a low-rank subspace regardless of the pre-training objective.

4.4 Layer-wise Task Affinity

Our layer-wise analysis reveals that the terminal layers of DINOv2 show a significant decline in explicitly decodable geometric information. To determine whether this reflects a loss of capacity or a shift toward semantic abstraction, we train identical linear probes for surface normal estimation, depth estimation, and semantic segmentation on NYUv2. Given that the input images and the frozen backbone remain constant, this controlled setup ensures that divergent layer-wise peak performances are driven by downstream task affinities. While this section primarily focuses on DINOv2, we extend this layer-wise analysis to MAE and iBOT. We observe a similar yet distinct trend: across all models, surface normal estimation consistently peaks in the earlier layers. However, the performance peaks for depth estimation and semantic segmentation do not exhibit a sharp

demarcation. Detailed results and comprehensive comparisons are provided in the Supplementary Material.

As illustrated in Fig. 7, the performance trajectories normalized against each task's layer-wise peak reveal a sequential transition between different tasks. Intermediate layers exhibit a strong affinity for local geometric features, with surface normal estimation peaking at layer 18, while depth estimation maintains structural accuracy longer and peaks at layer 21. This attenuation of explicit geometry is consistent with a steady improvement in semantic segmentation, which reaches its maximum at layer 24. Qualitative visualizations (Fig. 6) support this observation, showing that explicit geometric signals from intermediate layers are replaced by abstract representations in the terminal stages. This variation in task affinity suggests that relying solely on a terminal feature readout for multi-task dense prediction is sub-optimal, highlighting the need for depth-aware feature routing.

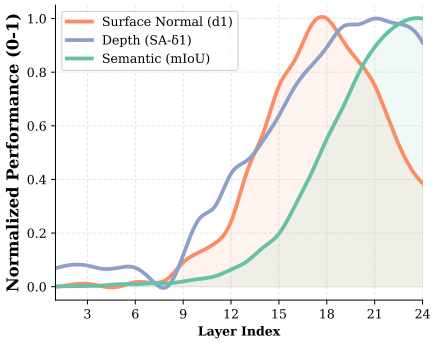


Fig. 7: Normalized probing performance. Geometric tasks peak and decline early, whereas high-level semantic abstraction peaks latest.

4.5 Robustness of the Extracted Subspaces

To rule out optimization artifacts, we evaluate the subspace similarity of the extracted bases ($\mathbf{V}_k$) across three random probe initializations. Fig. 8 reveals high consistency (*e.g.*, > 0.93 for DINOv2) at low ranks($k \leq 16$), suggesting that the geometric core is a stable structural property rather than an optimization byproduct. Conversely, the low similarity in the tail dimensions ($k \geq 64$) indicates a lack of stable geometric signals, making these bases highly susceptible to random initialization. It aligns with the low-rank compressibility shown in Fig. 2. Detailed stability metrics are provided in the Supplementary Material.

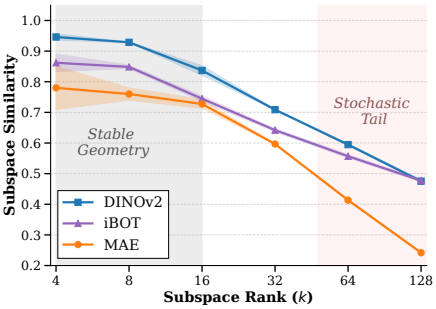


Fig. 8: Subspace Stability. Lines and color shaded regions denote the mean and standard deviation of subspace similarity across three random seeds over ranks k. High consistency at low ranks ($k \leq 16$) rules out optimization artifacts.

5 Related Work

5.1 Evolution of Probing Methodologies

Self-supervised ViT models naturally learn robust semantic and geometric representations that transfer effectively to dense prediction tasks [4, 12, 20, 31, 37]. While the broader advantages and limitations of these learning mechanisms have been systematically reviewed [26], standard probing efforts predominantly evaluate these models from a macroscopic perspective. For instance, recent studies confirm DINOv2's excellence in 2.5D view-centric geometry while revealing limitations in multi-view reasoning [3, 8, 13]. Crucially, these performance-driven evaluations treat the feature manifold as a black box, leaving the underlying organization of geometric primitives unexplained.

To address this limitation, recent research has shifted toward mechanistic interpretability. It is now established that visual concepts and spatial reasoning are not merely diffuse patterns but can be isolated into specific linear directions, attention heads [7], or dissected via overcomplete dictionary learning [16]. Our work extends this trajectory through controlled subspace intervention, directly isolating low-rank coordinate systems that encode explicit geometry.

5.2 Subspace Analysis and Intrinsic Dimensionality

The manifold hypothesis posits that high-dimensional visual data inherently resides on low-dimensional submanifolds [5, 36]. Empirical evidence confirms that highly parameterized pre-trained models exhibit remarkably low intrinsic dimensionality, yielding generalization bounds independent of their total parameter count [1, 28]. This pervasive low-rank bias in deep networks causes increasing depth to drive representational compression, strictly bounding the number of non-negligible singular values [17, 34]. In deep discriminative models, within-class variability of terminal-layer training activations collapses toward zero [32]. Specifically, Vision Transformers can experience exponential rank collapse within attention mechanisms, driving tokens toward a uniform state [14].

Crucially, self-supervised learning (SSL) objectives govern the topology of these low-rank manifolds. Contrastive and self-distillation frameworks explicitly constrain representations onto a low-dimensional, hard-shell hyperspherical manifold [27], effectively prioritizing invariant semantic concept extraction at the cost of spatial rigidity. In contrast, MAEs must preserve high-dimensional variance and high-frequency geometric primitives to satisfy dense pixel-level reconstruction constraints [21, 46]. Hybrid approaches combine these paradigms to balance global instance discriminability with fine-grained local awareness [23, 47]. Our analysis bridges these observations, revealing how distinct pre-training objectives dictate feature encoding formats and govern the layer-wise routing of explicit geometry and semantic abstraction.

6 Conclusion

This study explores the evolution of geometric representations in self-supervised vision models. Through targeted subspace intervention, we find that models like DINOv2 align geometric primitives into highly compressible, low-rank formats, whereas MAE disperses these signals across broader dimensions. Furthermore, the layer-wise analysis reveals a task affinity within deep representation spaces. We show that the decline of explicit geometric capacity in terminal layers is consistent with the emergence of peak semantic abstraction. By connecting these layer-wise representational shifts with downstream decoding complexity, our findings clarify the limitations of relying exclusively on terminal features for dense prediction. These findings provide a useful foundation for feature routing, more effective feature selection, and lightweight decoder design.

Limitations and Future Work. While our linear intervention rigorously isolates explicit geometric structures, it struggles to unroll highly non-linear feature entanglements. Furthermore, strict in-domain evaluation requires datasets with perfectly aligned depth, normals, and semantics. The scarcity of such comprehensive data limits broader outdoor or cross-domain validation. Additionally, due to resource constraints, we could not evaluate larger-scale models like ViT-G. Future work will explore non-linear topological interventions and synthetic multi-modal datasets to address these constraints.

Acknowledgements

This work was supported by JST CREST, Japan, under Grant Number JPMJCR2554, and by JSPS KAKENHI under Grant Number JP26K02785.

References

1. Aghajanyan, A., Zettlemoyer, L., Gupta, S.: Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In: Annual Meeting of the Association for Computational Linguistics (2020)
2. Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644 (2016)
3. Alam, N., Murali, L.K., Bharadwaj, S., Liu, P., Chung, T., Sharma, D., Kiran, K., Tam, W., Vegesna, B.K.S., et al.: The spatial blindspot of vision-language models. arXiv preprint arXiv:2601.09954 (2026)
4. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: Deep vit features as dense visual descriptors. arXiv preprint arXiv:2112.05814 **2**(3), 4 (2021)
5. Ansuini, A., Laio, A., Macke, J.H., Zoccolan, D.: Intrinsic dimension of data representations in deep neural networks. In: Neural Information Processing Systems (2019)
6. Bae, G., Budvytis, I., Cipolla, R.: Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13137–13146 (2021)

7. Bahador, N.: Mechanistic interpretability of fine-tuned vision transformers on distorted images: Decoding attention head behavior for transparent and trustworthy ai. arXiv preprint arXiv:2503.18762 (2025)
8. Banani, M.E., Raj, A., Maninis, K.K., Kar, A., Li, Y., Rubinstein, M., Sun, D., Guibas, L.J., Johnson, J., Jampani, V.: Probing the 3d awareness of visual foundation models. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 21795–21806 (2024)
9. Belinkov, Y.: Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics **48**, 207–219 (2021)
10. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4008–4017 (2020)
11. Cao, J., Leng, H., Lischinski, D., Cohen-Or, D., Tu, C., Li, Y.: Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 7068–7077 (2021)
12. Caron, M., Touvron, H., Misra, I., J'egou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 9630–9640 (2021)
13. Chen, X., Marks, M., Cheng, Z.: Probing the mid-level vision capabilities of self-supervised learning. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 30095–30105 (2024)
14. Dong, Y., Cordonnier, J.B., Loukas, A.: Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In: International conference on machine learning. pp. 2793–2803. PMLR (2021)
15. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Neural Information Processing Systems (2014)
16. Fel, T., Wang, B., Lepori, M.A., Kowal, M., Lee, A., Balestriero, R., Joseph, S., Lubana, E.S., Konkle, T., Ba, D., et al.: Into the rabbit hull: From task-relevant concepts in dino to minkowski geometry. arXiv preprint arXiv:2510.08638 (2025)
17. Garrod, C., Keating, J.P.: The persistence of neural collapse despite low-rank bias. arXiv preprint arXiv:2410.23169 (2024)
18. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The international journal of robotics research **32**(11), 1231–1237 (2013)
19. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
20. Goyal, P., Mahajan, D.K., Gupta, A.K., Misra, I.: Scaling and benchmarking self-supervised visual representation learning. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 6390–6399 (2019)
21. He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 15979–15988 (2021)
22. Hewitt, J., Liang, P.: Designing and interpreting probes with control tasks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp). pp. 2733–2743 (2019)
23. Huang, Z., Jin, X., Lu, C., Hou, Q., Cheng, M.M., Fu, D., Shen, X., Feng, J.: Contrastive masked autoencoders are stronger vision learners. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**, 2506–2517 (2022)

24. Jampani, V., Maninis, K.K., Engelhardt, A., Karpur, A., Truong, K., Sargent, K., Popov, S., Araujo, A., Martin Brualla, R., Patel, K., et al.: Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. Advances in Neural Information Processing Systems **36**, 76061–76084 (2023)
25. Jevtić, A., Reich, C., Wimbauer, F., Hahn, O., Rupprecht, C., Roth, S., Cremers, D.: Feed-forward scenedino for unsupervised semantic scene completion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6784–6796 (2025)
26. Khan, A., Sohail, A., Fiaz, M., Hassan, M., Afridi, T.H., Marwat, S.U., Munir, F., Ali, S., Naseem, H., Zaheer, M.Z., et al.: A survey of the self supervised learning mechanisms for vision transformers. arXiv preprint arXiv:2408.17059 (2024)
27. Kumar, A., Patel, V.M.: Learning on the manifold: Unlocking standard diffusion transformers with representation encoders. arXiv preprint arXiv:2602.10099 (2026)
28. Li, C., Farkhoor, H., Liu, R., Yosinski, J.: Measuring the intrinsic dimension of objective landscapes. arXiv preprint arXiv:1804.08838 (2018)
29. Man, Y., Zheng, S., Bao, Z., Hebert, M., Gui, L., Wang, Y.X.: Lexicon3d: Probing visual foundation models for complex 3d scene understanding. Advances in Neural Information Processing Systems **37**, 76819–76847 (2024)
30. Mao, Y., Liu, J., Liu, X.: Stealing stable diffusion prior for robust monocular depth estimation. arXiv preprint arXiv:2403.05056 (2024)
31. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
32. Papyan, V., Han, X., Donoho, D.L.: Prevalence of neural collapse during the terminal phase of deep learning training. Proceedings of the National Academy of Sciences of the United States of America **117**, 24652 – 24663 (2020)
33. Park, N., Kim, W., Heo, B., Kim, T., Yun, S.: What do self-supervised vision transformers learn? arXiv preprint arXiv:2305.00729 (2023)
34. Patel, N., Shwartz-Ziv, R.: Learning to compress: Local rank and information compression in deep neural networks. arXiv preprint arXiv:2410.07687 (2024)
35. Pimentel, T., Valvoda, J., Maudslay, R.H., Zmigrod, R., Williams, A., Cotterell, R.: Information-theoretic probing for linguistic structure. In: Annual Meeting of the Association for Computational Linguistics (2020)
36. Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., Goldstein, T.: The intrinsic dimension of images and its impact on learning. arXiv preprint arXiv:2104.08894 (2021)
37. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 12159–12168 (2021)
38. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European Conference on Computer Vision (2012)
39. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotný, D.: Vggt: Visual geometry grounded transformer. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5294–5306 (2025)
40. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International conference on machine learning. pp. 9929–9939. PMLR (2020)
41. Xie, Z., Geng, Z., Hu, J., Zhang, Z., Hu, H., Cao, Y.: Revealing the dark secrets of masked image modeling. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 14475–14485 (2022)

42. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10371–10381 (2024)
43. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
44. Zhan, G., Zheng, C., Xie, W., Zisserman, A.: A general protocol to probe large vision models for 3d physical understanding. Advances in Neural Information Processing Systems 37 (2023)
45. Zhang, J., Herrmann, C., Hur, J., Chen, E., Jampani, V., Sun, D., Yang, M.H.: Telling left from right: Identifying geometry-aware semantic correspondence. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3076–3085 (2023)
46. Zhang, Q., Wang, Y., Wang, Y.: How mask matters: Towards theoretical understandings of masked autoencoders. Advances in Neural Information Processing Systems **35**, 27127–27139 (2022)
47. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)