LEARNING TO PERCEIVE "WHERE": SPATIAL PRETEXT TASKS FOR ROBUST SELF-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing self-supervised learning (SSL) methods primarily learn object-invariant representations but often neglect the spatial structure and relationships among object parts. To address this limitation, we introduce Spatial Prediction (SP), a spatially aware pretext regression task that predicts the relative position and scale between a pair of disentangled local views from the same image. By modeling part-to-part relationships in a continuous geometric space, SP encourages representations to capture fine-grained spatial dependencies beyond invariant categorical semantics, thereby learning the compositional structure of visual scenes. SP is implemented as a decoupled plug-in and can be seamlessly integrated into diverse SSL frameworks. Extensive experiments show consistent improvements across image recognition, fine-grained classification, semantic segmentation, and depth estimation, as well as substantial gains in out-of-distribution robustness for object recognition. To evaluate spatial reasoning, we introduce (1) a position and scale prediction task on image patch pairs and (2) a jigsaw understanding task requiring patch reordering and recognition after reconstruction. Strong performance on these tasks indicates improved spatial structure and geometric awareness. Overall, explicitly modeling spatial information provides an effective inductive bias for SSL, leading to more structured representations and better generalization. Code and models will be released.

1 INTRODUCTION

Humans exhibit a strong capacity for spatial perception, enabling them to infer the relative positions, scales, and arrangements of objects and their parts in complex scenes Biederman (1987). Reasoning about part-to-part relationships is fundamental to understanding scene structure and visual composition, and underpins diverse real-world tasks, including object detection Wu et al. (2023); Wang et al. (2024), scene understanding Zhang et al. (2020); Bomatter et al. (2021); Liu et al. (2022); Jia et al. (2025); Khandelwal et al. (2023), instance segmentation Wang et al. (2023b); Han et al. (2024), 3D reconstruction Zhang et al. (2025), depth estimation Cai et al. (2025), and visual navigation Piriyajitakonkij et al. (2024); Yang et al. (2019); Liu et al. (2023); Zhang et al. (2018). However, modern self-supervised learning (SSL) methods Doersch et al. (2015); Assran et al. (2023); Chen et al. (2021); Caron et al. (2021); He et al. (2022); Xie et al. (2022); Chen et al. (2020); Oquab et al. (2024); Oord et al. (2018) primarily emphasize semantic invariance or local reconstruction, while largely overlooking explicit modeling of spatial relationships between image regions. As a result, the geometric structure of visual scenes remains under-constrained.

This limitation stems from the absence of explicit supervision for spatial reasoning in the SSL literature. Invariance-based approaches, including contrastive and self-distillation methods Chen et al. (2021); Caron et al. (2021); Chen & He (2021), enforce consistency across augmented views, promoting robustness to transformations such as cropping and scaling. However, by treating these transformations as nuisances, they suppress spatial variation and reduce sensitivity to relative position and scale O Pinheiro et al. (2020); Yun et al. (2022); Pariza et al. (2024); Xie et al. (2021). In parallel, reconstruction-based methods He et al. (2022); Xie et al. (2022); Bao et al. (2022); Assran et al. (2022) recover masked regions and capture fine-grained appearance statistics, but they operate at the patch level without spatial grounding. Consequently, learned representations exhibit limited sensitivity to part-to-part geometry Xie et al. (2023).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

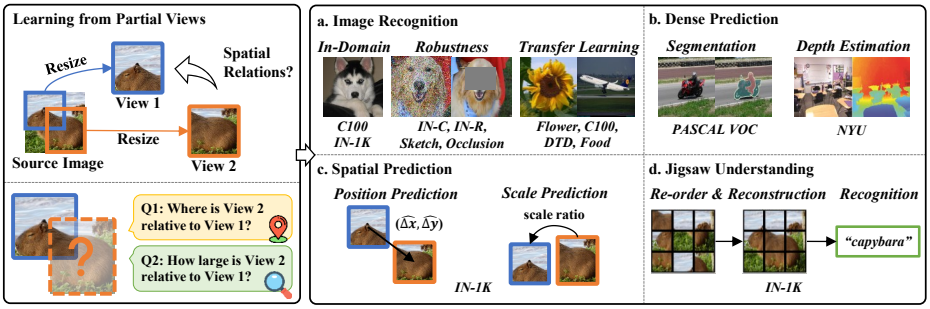

Figure 1: **Problem setting and task overview for visual representation learning from partial observations.** The left panel illustrates the problem setting: two cropped and resized views (orange and blue) are sampled from the same image, and the SSL model predicts the relative position and scale of the orange patch with respect to the blue reference. The right panel summarizes the evaluation tasks. (a) Image recognition includes in-domain classification on C100 Krizhevsky et al. (2009) and IN-1K Deng et al. (2009), out-of-distribution robustness on IN-C Hendrycks & Dietterich (2019), IN-R Hendrycks et al. (2021), Sketch Wang et al. (2019), and Occlusion, and cross-dataset transfer learning on Flowers Nilsback & Zisserman (2008), C100, DTD Cimpoi et al. (2014), and Food Bossard et al. (2014). (b) Dense prediction tasks include semantic segmentation on PASCAL VOC Everingham et al. (2010) and depth estimation on NYU Nathan Silberman & Fergus (2012). (c) Spatial prediction evaluates the ability of SSL models to estimate relative position and scale between two local views. (d) Jigsaw understanding evaluates patch reordering, reconstruction, and subsequent recognition. (c) and (d) correspond to our newly proposed benchmarks and tasks on IN-1K.

To close this gap, we propose **Spatial Prediction (SP)**, a spatially-aware pretext task that models geometric relationships between pairs of local regions via regression in continuous space. As illustrated in **Fig. 1 (left)**, given two independently sampled views from the same image, the SSL model predicts their relative position and scale. By turning spatial relations into explicit regression objectives, SP encourages models to capture how local regions are organized within a scene, rather than discarding geometry as invariance or restricting representations to discrete grid-aligned feature tokens. SP is an architecture-agnostic plug-in that integrates seamlessly into existing SSL frameworks. It introduces an auxiliary spatial reasoning branch without modifying the original training objective, enhancing geometric sensitivity while preserving semantic robustness. The resulting representations are both invariant to appearance changes and sensitive to spatial variation, providing a more balanced inductive bias for visual representation learning.

We evaluate learned representations of SSL models from two complementary perspectives (**Fig. 1, right**). First, we assess the effectiveness of models trained with SP on standard computer vision tasks, including image recognition, fine-grained classification, out-of-distribution recognition under occlusion and corruption (**Fig. 1a**), as well as image segmentation and depth estimation (**Fig. 1b**). To further analyze spatial reasoning, we establish two diagnostic benchmarks: (1) a position and scale prediction task, which measures the ability to infer relative geometry between pairs of image patches (**Fig. 1c**); and (2) a jigsaw understanding task, which evaluates whether models can recover disrupted spatial layouts and perform recognition after reconstruction (**Fig. 1d**). Together, these tasks directly probe spatial awareness beyond conventional SSL benchmarks. Key contributions are highlighted:

1. We introduce Spatial Prediction (SP), a spatially-aware pretext task that explicitly models geometric relationships between two local views by predicting their relative position and scale via regression. By formulating spatial relations as a direct training objective, SP goes beyond invariance- or reconstruction-driven SSL and encourages representations to capture fine-grained part-to-part dependencies and compositional scene structure.

2. We design SP as an architecture-agnostic plug-in that can be seamlessly integrated into diverse SSL frameworks, enabling spatial reasoning while preserving the semantic robustness of invariant representations.

3. We conduct extensive evaluations across five standard computer vision tasks, showing that SSL models trained with SP consistently outperform their counterparts. In addition, we introduce two spatial reasoning benchmarks: spatial prediction and jigsaw understanding. Experimental results provide direct evidence that SP improves spatial reasoning beyond standard benchmarks.

2 Related Works on Self-Supervised Learning in Vision

Self-supervised learning (SSL) learns visual representations from unlabeled data through pretext objectives. Early approaches explored relative patch prediction Doersch et al. (2015), jigsaw solving Noroozi & Favaro (2016), inpainting Pathak et al. (2016), colorization Zhang et al. (2016), and transformation prediction Gidaris et al. (2018).

Reconstruction-based methods. With Vision Transformers (ViTs) Dosovitskiy et al. (2020), masked image modeling has become a major SSL paradigm. Methods such as MAE He et al. (2022) and BEiT Bao et al. (2022) reconstruct masked inputs, while latent-space approaches such as data2vec and I-JEPA Baevski et al. (2022); Assran et al. (2023) predict representations in feature space. Despite their success, these methods can favor local or low-level reconstruction over holistic scene structure Xie et al. (2023); Balestriero & LeCun (2024) and generally lack explicit modeling of spatial relationships among image regions.

Discriminative SSL methods. Another major family learns representations by enforcing consistency across views of the same image, including contrastive learning He et al. (2020); Chen et al. (2020); Caron et al. (2020), self-distillation Grill et al. (2020); Caron et al. (2021), and clustering-based methods Caron et al. (2018); Asano et al. (2019). Extensions such as DenseCL Wang et al. (2021) preserve local correspondences, while iBOT Zhou et al. (2022) combines distillation with masked modeling. However, the emphasis on view invariance generally does not explicitly capture geometric relationships between regions and may suppress fine-grained spatial structure Wang et al. (2017); Lee et al. (2026).

Spatial and position-based SSL. Spatial pretext tasks date back to relative patch prediction Doersch et al. (2015) and jigsaw solving Noroozi & Favaro (2016), but these handcrafted objectives are not naturally integrated with modern ViT-based SSL. Recent methods incorporate spatial information more directly: I-JEPA Assran et al. (2023); Chen et al. (2024) conditions latent prediction on positional embeddings, while DropPos Wang et al. (2023a) reconstructs masked positional coordinates. Yet, position is primarily used as conditioning or a reconstruction target rather than to explicitly model geometric relations between views.

Concurrent with our work, PART Ayoughi et al. (2025) also predicts relative position and scale between views. Unlike PART's standalone pretext objective, our spatial prediction (SP) serves as a plug-in regularizer for existing SSL objectives, preserving semantic invariance while adding explicit geometric supervision. SP further employs rejection sampling to construct informative view pairs with balanced spatial distributions, avoiding excessive overlap, biased targets, and semantically uninformative patches. In contrast, PART does not explicitly control the view-pair sampling distribution. We further evaluate SP across diverse downstream tasks and dedicated spatial reasoning benchmarks, examining both semantic representation and spatial reasoning ability.

3 Spatial Prediction (SP): A Spatially-Aware Pretext Task

We propose a plug-and-play pretext task, termed **Spatial Prediction (SP)**, that augments SSL with explicit geometric supervision during pre-training (**Fig. 2**). It can be seamlessly integrated into diverse SSL frameworks without modifying their model architectures. SP leverages Vision Transformer (ViT) Dosovitskiy et al. (2020) feature tokens to model spatial dependencies. At a high level, SP shares the encoder $\psi(\cdot)$ with the original SSL objective and introduces an additional spatial prediction branch. Both objectives are jointly optimized during pre-training, enabling the model to learn complementary semantic and geometric representations. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, where H and W denote the height and width, the encoder $\psi(\cdot)$ outputs a sequence of patch tokens $\mathbf{Z} \in \mathbb{R}^{L\times D}$ and a classification token $\mathbf{z} \in \mathbb{R}^{1\times D}$, with L denoting the number of tokens and D the feature dimension. SP operates on $\mathbf{Z}$ and $\mathbf{z}$, as described below.

3.1 Curating Pairs of Augmented Local Views for Pre-Texting Tasks

To train SSL models with SP, we construct pairs of augmented local views from the same image. Given a source image $\mathbf{I}$, we first apply standard augmentations (e.g., random horizontal flipping), followed by rejection sampling to obtain two local views. The sampling is guided by three criteria: (i) views should be sufficiently local to capture partial object information, but not so small that they lose

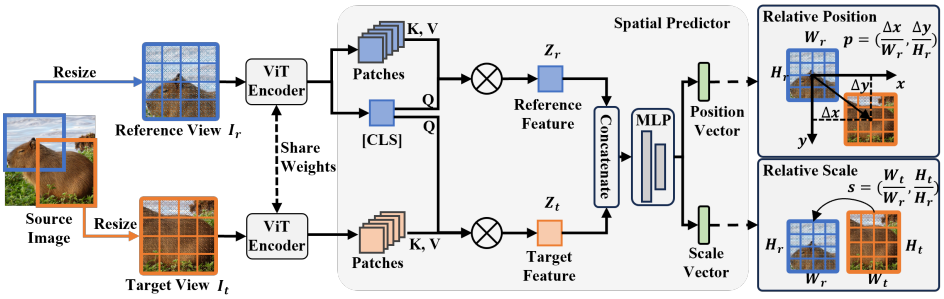

Figure 2: **Overview of Spatial Prediction (SP).** Given a reference view I_r and a target view I_t sampled from the same image, both views are encoded by a shared Vision Transformer (ViT) Dosovitskiy et al. (2020). The class token ([CLS]) of the reference view serves as the query (Q), while patch tokens from both reference and target views act as keys (K) and values (V) to compute a cross-attention-like interaction ($\otimes$), producing reference and target features. This operation is parameter-free, as it directly uses token embeddings without additional projections for Q, K, and V. The resulting features $\mathbf{Z}_r$ and $\mathbf{Z}_t$ are concatenated and fed into a two-layer MLP to regress the relative position and scale, supervised using ℓ_2 loss against ground-truth targets. As illustrated in the right panels, the ground-truth spatial relationship is defined by the relative offset $(\Delta x, \Delta y)$ from the reference view to the target view, along with the relative scale normalized by the reference view dimensions (H_r, W_r), where (H_r, W_r) and (H_t, W_t) denote the height and width of the reference and target views, respectively.

semantic content; (ii) the relative position and scale targets are sampled to be approximately uniform, avoiding skewed or long-tailed distributions that could bias regression toward trivial solutions; and (iii) a non-trivial spatial displacement is enforced to prevent large overlaps that would make the prediction task overly easy due to shortcut cues. We validate the distributions of ground truth relative position and scale after rejection sampling in **Fig.S2**.

To further mitigate shortcut solutions, we resize both sampled local views to a fixed size and apply independent color augmentations. We avoid extra spatial transformations (e.g., rotation or further cropping) that would distort the ground-truth relative positions and scales. These design choices discourage the SSL models from over-reliance on low-level cues such as color or template matching, and instead promote semantic correspondence for spatial reasoning.

Without loss of generality, we denote the first local view as the reference view I_r and the second as the target view I_t, with corresponding heights and widths (H_r, W_r) and (H_t, W_t), respectively. Since I_r and I_t may differ in size and location, we define the ground-truth relative position p and scale s normalized by the reference view dimensions (H_r, W_r) (see **Fig. 2**). This normalization is a key design choice, as it makes the geometric ground truth scale-invariant and independent of the absolute image size of I_r, allowing SSL models to focus on relative spatial structure.

3.2 Spatial Representation Learning

The encoder $\psi(\cdot)$ takes the reference and target views I_r and I_t as inputs and produces their patch tokens and classification tokens, where $Z_r, Z_t \in \mathbb{R}^{L\times D}$ and $z_r, z_t \in \mathbb{R}^{1\times D}$:

$$[Z_r, z_r] = \psi(I_r), \quad [Z_t, z_t] = \psi(I_t). \tag{1}$$

Motivated by human spatial reasoning, we formulate the task as reference-conditioned spatial inference, where a reference view provides an anchor for estimating the relative position and scale of a target view. The model first establishes a reference coordinate system from I_r and then performs spatial reasoning over I_t conditioned on this reference. Thus, we introduce a cross-attention-based mechanism to model this interaction. Specifically, z_r serves as a query, while Z_r and Z_t serve as keys and values for retrieving reference-aware and target-aware features:

$$h_r = \text{Softmax}\big(\text{norm}(z_r)Z_r^\top\big)Z_r, \quad h_t = \text{Softmax}\big(\text{norm}(z_r)Z_t^\top\big)Z_t, \tag{2}$$

where $\top$ denotes matrix transpose and $\text{norm}(\cdot)$ is layer normalization for training stability. Unlike standard cross-attention, we directly use patch tokens and classification tokens without additional

linear projections for queries, keys, and values. This design enforces stronger supervision on token interactions for spatial reasoning while avoiding additional learnable parameters.

The retrieved features are then concatenated to form a joint representation. A spatial predictor $\theta(\cdot)$, implemented as a two-layer MLP with hidden dimension 384 and ReLU activation, regresses the spatial outputs:

$$[\hat{p}, \hat{s}] = \theta([h_r; h_t]), \tag{3}$$

where $\hat{p} \in \mathbb{R}^2$ and $\hat{s} \in \mathbb{R}^2$ denote the predicted relative spatial displacement and scale, respectively.

3.3 Training Objectives and Implementation Details

The proposed SP objective is introduced as an auxiliary loss in addition to the base SSL objective, denoted as $\mathcal{L}_{\text{base}}$ (e.g., contrastive loss for MoCo v3 (Chen et al., 2021), distillation loss for DINO Caron et al. (2021), or reconstruction loss for MAE He et al. (2022)). Detailed schematics of SP integration into each SSL framework are provided in **Fig. S1**.

The spatial predictor is trained to regress the relative spatial configuration between two views using ℓ_2 losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \mathcal{L}_{\text{SP}}, \quad \text{where} \quad \mathcal{L}_{\text{SP}} = \lambda_p \|\hat{\mathbf{p}} - \mathbf{p}\|_2^2 + \lambda_s \|\hat{\mathbf{s}} - \mathbf{s}\|_2^2, \tag{4}$$

$\lambda_p = 0.1$ and $\lambda_s = 0.1$ balance the two regression terms. See **Tab. S2** for hyperparameter analysis.

We pre-train all SSL models with SP from scratch. All experiments are conducted on two NVIDIA RTX A6000 GPUs with a batch size of 256. Optimization uses AdamW with a cosine learning-rate schedule. All hyperparameters follow the default configurations of the original SSL implementations without SP. See **Tab. S1** for a summary of pre-training hyperparameters for all SSL models.

4 Experiments

We introduce a comprehensive evaluation suite covering 7 downstream tasks, 11 datasets, 6 SSL models, 2 backbones, and 7 metrics.

4.1 Datasets and Metrics

Datasets for pre-training SSL models: CIFAR-100 (C100) Krizhevsky et al. (2009), a small-scale dataset with 60K images of size 32×32 across 100 classes, and **ImageNet-1K (IN-1K)** Deng et al. (2009), a large-scale dataset with 1.28M images of size 224×224 across 1,000 classes. For in-domain image classification, we evaluate on the corresponding test sets. For all subsequent tasks, we use SSL backbones pretrained on IN-1K. We report Top-1 classification accuracy (Acc in %) of SSL models.

Datasets for robustness tests in image classification: We evaluate SSL model robustness using several ImageNet-derived benchmarks. **ImageNet-C (IN-C)** Hendrycks & Dietterich (2019) measures robustness to common corruptions (e.g., noise, blur, weather) across multiple severity levels. **ImageNet-R (IN-R)** Hendrycks et al. (2021) and **ImageNet-Sketch (Skt)** Wang et al. (2019) assess robustness to domain and appearance shifts, including artistic renditions and sketch-based images that emphasize shape over texture and color. **ImageNet-Occlusion (Occ)** is our synthetic benchmark, where we systematically apply occlusions by overlaying opaque rectangular masks at random locations with varying coverage ratios of 0.1, enabling controlled evaluation of robustness to partial visibility and spatial reasoning under missing visual information. We report Top-1 classification accuracy (Acc. in %) of SSL models across all datasets, except IN-C, where we report mean Corruption Error (mCE), computed as the average normalized error over all corruption types and severity levels.

Datasets for transfer learning in fine-grained classification: We evaluate transfer learning ability of SSL models on diverse fine-grained datasets. **Flowers-102 (Flower)** Nilsback & Zisserman (2008) contains 102 flower categories with subtle inter-class differences. **DTD** Cimpoi et al. (2014) focuses on texture recognition across 47 categories of materials and surface patterns. **Food-101 (Food)** Bossard et al. (2014) includes 101 food categories with high visual variability. We report Top-1 classification accuracy (Acc in %) of SSL models.

Dataset for transfer learning in semantic segmentation: We evaluate semantic segmentation performance of SSL models. **PASCAL VOC (VOC)** Everingham et al. (2010) is used for semantic segmentation, with pixel-level annotations on images of size 512×512 across 21 categories. We report mean Intersection over Union (mIoU), computed as the average intersection-over-union between predicted and ground-truth segmentation masks across classes.

Dataset for transfer learning in depth estimation: We evaluate depth estimation performance of SSL models. **NYU v2 (NYU)** Nathan Silberman & Fergus (2012) is used for depth estimation, providing aligned RGB images and depth maps for indoor scenes. We report root mean squared error (RMSE), computed as the square root of the mean squared difference between predicted and ground-truth depth values.

Dataset for spatial prediction: We use images from the IN-1K test set and sample pairs of local patches from the same image following the pre-training protocol (**Sec.3**). Given a reference patch, the model is required to predict the relative position and scale of the target patch. We report L2 distance between the predicted and ground-truth relative position and scale.

Dataset for jigsaw understanding: We use images from the IN-1K test set, partition each image into a 3×3 grid, and shuffle the patches using a predefined set of 1,000 permutations. Models are required to predict the permutation index (1,000-way classification), reconstruct the image, and perform recognition on the restored input. We report Top-1 accuracy (Acc in %), measuring (i) permutation classification accuracy and (ii) object recognition accuracy on reconstructed images.

4.2 Baselines, Backbones, and Experimental Protocols

We evaluate three representative SSL frameworks: **MoCo v3** (Chen et al., 2021), **DINO** (Caron et al., 2021), and **MAE** (He et al., 2022), covering contrastive learning, self-distillation, and reconstruction-based pretraining. We follow their original data augmentation, pre-training, linear probing, and fine-tuning protocols (Chen et al., 2021; Caron et al., 2021; He et al., 2022). For SSL methods integrated with our SP module, we retain their default pre-training, linear probing, and fine-tuning configurations, consistent with their standard counterparts without SP (He et al., 2022; Caron et al., 2021; Oquab et al., 2024; Assran et al., 2023). In linear probing, the backbone is frozen and only a linear classifier is trained end-to-end. The learning rate is selected via grid search based on validation performance.

Under a fixed computational budget, we use ViT-S/16 (Dosovitskiy et al., 2020) for MoCo v3 and DINO, and ViT-B/16 for MAE. We note that using different backbone sizes does not affect fairness of comparison, as we focus on within-model comparisons with and without SP.

For in-domain image classification, fine-grained transfer learning, and depth estimation, we report linear probing results. For robustness evaluation, we directly apply the same linear probes to the target datasets without additional fine-tuning. For semantic segmentation, we report fine-tuning results.

For the spatial prediction task, we use the same cross-attention-based mechanism as in spatial representation learning to predict the relative position and scale between two local views (**Sec. 3.2**).

For the jigsaw understanding task, we independently encode 9 shuffled image patches using the SSL backbone and extract their classification tokens. These classification tokens serve as queries that attend over all patch tokens (keys and values) from all patches via a cross-attention mechanism, as described in **Fig. S4**. The resulting outputs from the 9 cross-attention operations are concatenated and passed through a 2-layer MLP to predict the permutation index. For object recognition on reconstructed images, we reuse the same linear probes trained for in-domain classification without further fine-tuning.

5 Results

5.1 SP Excels at Spatial Reasoning While Preserving Robust Semantic Representations.

SP does not impair, but instead enhances semantic representation learning. As shown in **Table 1a**, SSL models with SP achieve consistent gains over their baselines on both CIFAR-100 and IN-1K for in-domain image classification. The improvements are particularly pronounced for MAE and

Table 1: **Performance comparison across all downstream tasks except spatial reasoning.** Each column corresponds to a task, together with its dataset and evaluation metric. Rows are grouped in pairs for SSL models with and without SP, where each pair shares the same backbone. All reported Acc values denote Top-1 classification accuracy (%). For IN-C, VOC, and NYU, we report mCE↓, mIoU↑, and RMSE (scaled by $\times 10^2$)↓, respectively. See **Sec. 4** for experimental details. Each experiment is repeated three times, and we report the mean performance with standard deviations in ($\pm$ std). Best results are shown in bold.

	Image Recognition										Dense Prediction	
	a. In-Domain		b. Robustness				c. Transfer Learning				d. Seg.	e. Depth
	C100* Acc↑	IN-1K Acc↑	IN-C mCE↓	IN-R Acc↑	Skt Acc↑	Occ Acc↑	Flw Acc↑	C100 Acc↑	DTD Acc↑	Food Acc↑	VOC mIoU↑	NYU RMSE↓
MAE	39.7 ± 0.5	44.6 ± 0.6	108.5 ± 0.3	7.2 ± 0.3	5.5 ± 0.3	36.7 ± 0.3	72.7 ± 0.2	65.6 ± 0.1	55.5 ± 0.2	56.8 ± 0.2	51.0 ± 0.1	58.2 ± 0.2
+SP	**43.5 ± 0.4**	**52.6 ± 0.4**	**104.8 ± 0.2**	**9.2 ± 0.5**	**7.6 ± 0.4**	**41.6 ± 0.9**	**80.1 ± 0.2**	**67.6 ± 0.2**	**56.5 ± 0.2**	**60.7 ± 0.2**	**51.2 ± 0.2**	**57.3 ± 0.3**
MoCo v3	62.1 ± 0.3	61.1 ± 0.6	85.6 ± 0.2	13.9 ± 0.2	13.6 ± 0.1	57.3 ± 0.3	63.7 ± 0.4	62.2 ± 0.1	57.2 ± 0.3	58.5 ± 0.1	58.5 ± 0.1	60.1 ± 0.7
+SP	**64.6 ± 0.3**	**66.1 ± 0.6**	**81.4 ± 0.2**	**16.7 ± 0.1**	**15.9 ± 0.1**	**60.5 ± 0.9**	**88.7 ± 1.1**	**77.2 ± 0.1**	**66.7 ± 0.3**	**72.8 ± 0.2**	**60.2 ± 0.1**	**59.9 ± 0.6**
DINO	50.5 ± 0.2	72.8 ± 0.2	**86.3 ± 0.2**	**15.7 ± 0.1**	**14.6 ± 0.1**	60.1 ± 0.3	86.4 ± 0.5	75.8 ± 0.1	**66.5 ± 0.2**	**70.5 ± 0.2**	46.4 ± 0.1	57.3 ± 0.9
+SP	**53.0 ± 0.4**	**73.0 ± 0.2**	88.5 ± 0.3	**15.7 ± 0.1**	**14.6 ± 0.1**	**60.5 ± 0.7**	**87.3 ± 0.2**	**76.1 ± 0.2**	65.7 ± 0.1	70.3 ± 0.2	**47.2 ± 0.1**	**55.8 ± 0.3**

MoCo v3, while even the strong DINO benefits from SP. These results indicate that explicit spatial supervision does not hinder semantic representation learning; instead, it enhances the semantic quality of learned representations. Further comparisons are shown in **Table S3**. Using the same additional SP crops without relation modeling yields smaller gains, showing that the improvement comes from modeling spatial relations rather than merely seeing more crops. MoCo v3 + SP also outperforms PART Ayoughi et al. (2025), which directly learns from spatial supervision, supporting our design of SP as a plug-in that complements the original SSL objective.

To further analyze model behavior in in-domain classification, we visualize 2D attention maps on example images in **Fig. 3**, illustrating where SSL models attend during classification. Without SP, attention maps (e.g., a sitting dog in Row 1) are often diffuse and partially influenced by background context. In contrast, models with SP focus more on semantically meaningful object regions. While DINO already highlights the dog face, DINO with SP yields sharper localization and more comprehensive coverage of the dog body. This behavior suggests that SP encourages object-centric representations, which further improves the quality of learned semantic features.

SP improves robustness to noise, occlusion and corruptions and reduces bias towards textures. We evaluate robustness on IN-C, IN-R, Skt, and Occ without fine-tuning. As shown in **Tab. 1b**, SP consistently improves performance for MAE and MoCo v3 across all benchmarks. Gains are particularly notable on Skt and IN-R, where texture cues are suppressed and shape information dominates; e.g., MoCo v3 + SP improves by +2.3% on Skt and +2.8% on IN-R. This aligns with prior findings that SSL models without SP tend to over-rely on texture cues Cai et al. (2025). On IN-C, SP reduces mCE across all backbones (e.g., 108.5→104.8 for MAE and 85.6→81.4 for MoCo v3), indicating improved robustness to noise and blur. These results suggest that SP introduces structural inductive bias, encouraging reliance on global object layout when local cues are degraded.

SP enhances feature transferability, especially for fine-grained recognition. We evaluate transfer learning on Flw, Food, DTD, and C100. As shown in **Tab. 1c**, SP consistently improves performance across datasets. Gains are large on fine-grained tasks; e.g., MoCo v3 + SP improves by 25% on Flw, highlighting the benefit of part-level spatial modeling for subtle part-to-whole distinctions.

SP provides spatially grounded features for semantic segmentation. We evaluate semantic segmentation on VOC. As shown in **Tab. 1d**, SP consistently improves mIoU across all SSL methods. These results suggest that SP enhances spatial grounding in learned features, benefiting pixel-level prediction. By preserving localized semantic structures rather than collapsing representations into a global descriptor, SP enables better exploitation of spatial information during fine-tuning, leading to improved object boundaries and precise localization.

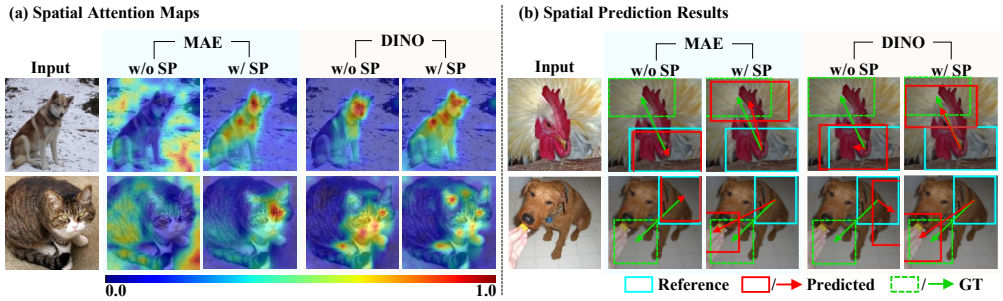

Figure 3: **Example visualizations of learned representations by SSL models with and without SP. (a) Qualitative comparison of spatial attention maps.** Following Caron et al. (2021), we visualize attention by computing the attention weights between the classification token z and all patch tokens Z, reshaping them into a 2D spatial map, and upsampling to the input image size. Each row corresponds to an input image, and columns show attention maps overlaid on the images for MAE and DINO, with and without SP. Heatmaps are normalized to [0,1]; see the colorbar for scale. **(b) Qualitative comparison of spatial reasoning results.** Predicted relative positions and scales between two local views are visualized on the original image. Each row corresponds to an input image, and columns show results for MAE and DINO, with and without SP. Blue boxes denote reference views, red boxes denote predicted positions and scales, and green dashed boxes denote ground truth; arrows indicate predicted (red) and ground-truth (green) displacement vectors. See **Fig. S3** for examples on MoCo v3.

SP enables representations that capture geometric structure for depth estimation. We evaluate depth estimation on NYU. As shown in **Tab. 1e**, SP consistently reduces RMSE across all backbones, with the largest improvement for DINO (-1.5 RMSE). These results suggest that SP promotes representations that capture 3D geometric information. By explicitly modeling spatial relationships during pre-training, SP encourages the use of geometric cues such as depth ordering and surface continuity for spatial inference.

SP provides stronger spatial supervision than implicit positional embeddings in standard SSL. We evaluate spatial prediction via relative position and scale estimation. As shown in **Tab. 2a–b**, incorporating SP consistently reduces L2 errors for both position and scale across all SSL backbones. These results indicate that SP provides explicit geometric supervision, enabling representations to encode precise spatial relationships beyond implicit positional embeddings used in standard SSL.

Visualization results in **Fig. 3b** further support this observation. Standard SSL models often exhibit noticeable errors in both the direction and magnitude of relative position offsets, reflected by misaligned displacement vectors and bounding boxes. In contrast, models with SP produce predictions whose spatial predictions (red boxes) closely match the ground truth (green dashed boxes), with displacement vectors well aligned in both direction and magnitude.

The benefits of SP in spatial reasoning transfer to downstream tasks via coherent part-to-whole reconstruction. We demonstrate the effectiveness of SP for spatial reasoning using the jigsaw understanding task. As shown in **Tab. 2c**, SP consistently improves permutation prediction accuracy across all SSL backbones, with large gains for MoCo v3 (69.87 $\rightarrow$ 90.24) and MAE (77.95 $\rightarrow$ 98.58). This indicates that SP significantly strengthens the ability to infer global spatial configurations from disjoint local patches, a capability that is limited in standard SSL models. This improved spatial reasoning further benefits downstream tasks such as recognition on reconstructed images, where patches are rearranged according to predicted permutations. As shown in **Tab. 2d**, SP consistently improves recognition accuracy across all models, including MAE (39.19 $\rightarrow$ 48.21) and MoCo v3 (57.09 $\rightarrow$ 64.23). These results suggest that the recovered spatial structures are both geometrically accurate and semantically meaningful. By enforcing spatial consistency during pre-training, SP encourages representations that preserve coherent part-to-whole relationships under spatial reorganization.

5.2 Ablation Studies Reveal Key Design Choices in SP.

We conduct ablation studies on the SP design using a MoCo v3 (ViT-Tiny) backbone pre-trained on CIFAR-100 for in-domain image classification. Top-1 accuracy is reported in **Tab. 3**.

Table 2: **Performance comparison in spatial prediction and jigsaw understanding tasks.** Columns are grouped into spatial prediction (position and scale; evaluated by L2 distance ↓) and jigsaw understanding (permutation order prediction and recognition on reconstructed images; evaluated by Top-1 accuracy ↑ (%)). Rows are grouped in pairs, each corresponding to a SSL method with and w/o our SP. Best in bold.

	Spatial Prediction		Jigsaw Understanding	
	a. Position	b. Scale	c. Order	d. Recog.
MAE	0.92	0.39	77.95	39.19
+SP	**0.61**	**0.35**	**98.58**	**48.21**
MoCo v3	1.47	0.45	69.87	57.09
+SP	**1.16**	**0.41**	**90.24**	**64.23**
DINO	1.32	0.43	88.45	63.45
+SP	**1.20**	**0.42**	**96.17**	**64.56**

Table 3: **Analysis of SP design choices.** L2 regression supervises relative position (Col1) and scale (Col2). We compare cross-attention mechanisms with (PAttn, Col3) and without (FAttn, Col4) linear projections for queries, keys, and values, where checkmarks (✓) indicate their presence. Experiments are conducted using MoCo v3 on C100 for in-domain image classification. Performance is reported as Top-1 accuracy (%)↑, with the best results in bold.

Variant	Spatial Supervision		Spatial Predictor		Acc
	Position	Scale	PAttn	FAttn	
1	✓			✓	63.9
2		✓		✓	63.0
3	✓	✓			64.3
4	✓	✓	✓		63.7
5 (Ours)	✓	✓		✓	**64.8**

Joint supervision of position and scale is critical for visual representation learning. In Columns 1 and 2 of **Tab. 3**, position-only L2 regression (Variant 1) achieves 63.9% accuracy, while scale-only supervision (Variant 2) yields 63.0%. However, both are outperformed by our full SP (Variant 5), which jointly models position and scale. This indicates that the two objectives provide complementary supervision, and that joint regression better captures spatial structure in a unified representation.

Projection-free cross-attention is effective for spatial representation learning. In Columns 3 and 4 of **Tab. 3**, we evaluate the role of cross-attention and its parameterization. Variant 3 removes cross-attention and directly concatenates patch tokens from the two views I_r and I_t, followed by a 2-layer MLP for prediction, achieving 64.3% accuracy. This simple concatenation of feature tokens already shows strong performance, suggesting that simple aggregation provides a strong baseline for spatial reasoning. Next, Variant 4 introduces linear projections for queries, keys, and values in cross-attention. This degrades performance to 63.7%, compared to 64.8% for our full SP. This suggests that additional parameterization may dilute the spatial supervision signal. In contrast, preserving raw token interactions while enforcing structured cross-view attention is more effective for learning spatial relationships.

6 Discussion

We present Spatial Prediction (SP), a spatially aware pretext task that explicitly models relative position and scale between two local views from the same image during SSL. SP serves as a simple plug-in for existing SSL frameworks, without modifying their architectures or incurring additional inference cost. We introduce a comprehensive evaluation suite spanning 7 downstream tasks, 11 datasets, 6 SSL models, 2 backbones, and 7 metrics. Among these, we furthur propose two spatial reasoning benchmarks: spatial prediction and jigsaw understanding. Empirical results show that existing SSL models exhibit limited spatial reasoning ability. In contrast, SP substantially improves spatial reasoning while enhancing semantic representation quality, robustness to corruptions and occlusions, and shape bias over texture bias. Moreover, SP-learned representations are spatially grounded and transfer effectively to semantic segmentation and fine-grained classification. They also capture 3D geometric structure, improving depth estimation. Our benchmarks further show that SP's inductive bias enables recovery of structured spatial layouts from disorganized patches, whereas implicit positional embeddings in standard SSL are insufficient for spatial reasoning. Overall, SP provides a simple yet effective mechanism for bridging semantic and geometric learning in self-supervised representations, motivating future work on spatial, temporal, and physical reasoning in visual foundation models. Despite its strong spatial reasoning performance, SP is currently limited to 2D space. Extending it to 3D or temporal constraints is an important direction for future work. As a plug-in regularizer, SP may also benefit from more advanced local-view sampling strategies. We hope stronger spatial reasoning capabilities will enable more capable vision foundation models for robotics, assistive technologies, and embodied AI.

AI USE STATEMENT

In this work, we used generative AI tools for polishing writing. We have not used generative AI tools for ideation or drafting sections of the paper, and generating synthetic datasets and proving mathematical claims are not applicable to this work. We have reviewed all AI-assisted work. We checked LLM-polished content and ensure the correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European conference on computer vision*, pp. 456–473. Springer, 2022.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, June 2023.

Melika Ayoughi, Samira Abnar, Chen Huang, Chris Sandino, Sayeri Lala, Eeshan Gunesh Dhekane, Dan Busbridge, Shuangfei Zhai, Vimal Thilak, Josh Susskind, et al. How parts assemble into wholes: Learning the relative composition of images. *arXiv preprint arXiv:2506.03682*, 2025.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1298–1312. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/baevski22a.html.

Randall Balestriero and Yann LeCun. Learning by reconstruction produces uninformative features for perception. *arXiv preprint arXiv:2402.11337*, 2024.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2022. URL https://arxiv.org/abs/2106.08254.

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

Philipp Bomatter, Mengmi Zhang, Dimitar Karev, Spandan Madan, Claire Tseng, and Gabriel Kreiman. When pigs fly: Contextual reasoning in synthetic and natural scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 255–264, 2021.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Yusen Cai, Bhargava Satya Nunna, Qing Lin, and Mengmi Zhang. Learning to see through a baby's eyes: Early visual diets enable robust visual intelligence in humans and machines. *arXiv preprint arXiv:2511.14440*, 2025.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. URL https://arxiv.org/abs/2104.14294.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132(1):208–223, 2024.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021. URL https://arxiv.org/abs/2104.02057.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Victor Guilherme Turrisi da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. solo-learn: A library of self-supervised methods for visual representation learning. *Journal of Machine Learning Research*, 23(56):1–6, 2022. URL http://jmlr.org/papers/v23/21-1155.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2018. URL https://arxiv.org/abs/1803.07728.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

Shuangpeng Han, Ziyu Wang, and Mengmi Zhang. Flow snapshot neurons in action: Deep neural networks generalize to biological motion perception. *Advances in Neural Information Processing Systems*, 37:53732–53763, 2024.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.

Yanhao Jia, Ji Xie, S Jivaganesh, Hao Li, Xu Wu, and Mengmi Zhang. Seeing sound, hearing sight: Uncovering modality bias and conflict of ai models in sound localization. *arXiv preprint arXiv:2505.11217*, 2025.

Naitik Khandelwal, Xiao Liu, and Mengmi Zhang. Adaptive visual scene understanding: incremental scene graph generation. *arXiv preprint arXiv:2310.01636*, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Joohyung Lee, Changhun Kim, Hyunsu Kim, Kwanhyung Lee, and Juho Lee. Soft equivariance regularization for invariant self-supervised learning. *arXiv preprint arXiv:2603.06693*, 2026.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

Xiao Liu, Ankur Sikarwar, Gabriel Kreiman, Zenglin Shi, and Mengmi Zhang. Reason from context with self-supervised learning. *arXiv preprint arXiv:2211.12817*, 2022.

Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 69–84, Cham, 2016. Springer International Publishing.

Pedro O O Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. *Advances in neural information processing systems*, 33:4489–4500, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL `https://arxiv.org/abs/2304.07193`.

Valentinos Pariza, Mohammadreza Salehi, Gertjan Burghouts, Francesco Locatello, and Yuki M Asano. Near, far: Patch-ordering enhances vision foundation models' scene understanding. *arXiv preprint arXiv:2408.11054*, 2024.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Maytus Piriyajitakonkij, Mingfei Sun, Mengmi Zhang, and Wei Pan. Tta-nav: Test-time adaptive reconstruction for point-goal navigation under visual corruptions. *arXiv preprint arXiv:2403.01977*, 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL `https://arxiv.org/abs/2508.10104`.

Igor Susmelj, Matthias Heller, Philipp Wirth, Jeremy Prescott, Malte Ebner, and et al. Lightly. URL `https://github.com/lightly-ai/lightly`.

Haochen Wang, Junsong Fan, Yuxi Wang, Kaiyou Song, Tong Wang, and ZHAO-XIANG ZHANG. Droppos: Pre-training vision transformers by reconstructing dropped positions. *Advances in Neural Information Processing Systems*, 36:46134–46151, 2023a.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pp. 10506–10518, 2019.

Xiaolong Wang, Kaiming He, and Abhinav Gupta. Transitive invariance for self-supervised visual representation learning. In *Proceedings of the IEEE international conference on computer vision*, pp. 1329–1338, 2017.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3024–3033, 2021.

Ziyu Wang, Mike Zheng Shou, and Mengmi Zhang. Object-centric learning with cyclic walks between parts and whole. *Advances in Neural Information Processing Systems*, 36:9388–9408, 2023b.

Ziyu Wang, Shuangpeng Han, and Mengmi Zhang. Pose prior learner: Unsupervised categorical prior learning for pose estimation. *arXiv preprint arXiv:2410.03858*, 2024.

Jay Zhangjie Wu, David Junhao Zhang, Wynne Hsu, Mengmi Zhang, and Mike Zheng Shou. Label-efficient online continual object detection in streaming video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19246–19255, 2023.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16684–16693, 2021.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9653–9663, 2022.

Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14475–14485, 2023.

Kaiyu Yang, Olga Russakovsky, and Jia Deng. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2051–2060, 2019.

Sukmin Yun, Hankook Lee, Jaehyung Kim, and Jinwoo Shin. Patch-level representation learning for self-supervised vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8354–8363, 2022.

Mengmi Zhang, Keng Teck Ma, Shih-Cheng Yen, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Egocentric spatial memory. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 137–144. IEEE, 2018.

Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12985–12994, 2020.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016.

Zhengquan Zhang, Feng Xu, and Mengmi Zhang. Peering into the unknown: Active view selection with neural uncertainty maps for 3d reconstruction. *arXiv preprint arXiv:2506.14856*, 2025.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer, 2022. URL https://arxiv.org/abs/2111.07832.

SUPPLEMENTARY MATERIALS

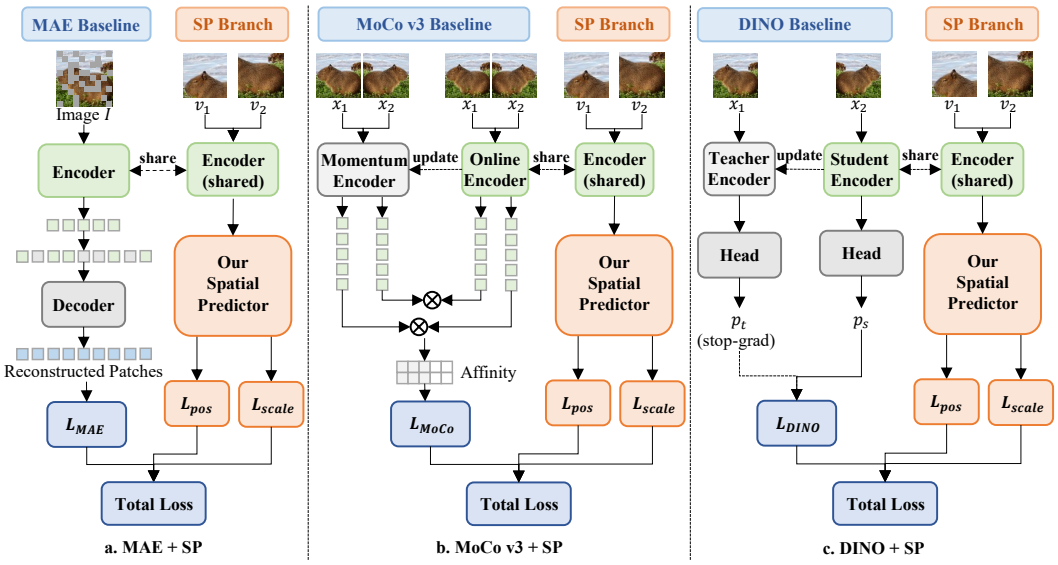

Figure S1: **Detailed architecture of Spatial Prediction (SP) integration across diverse SSL frameworks.** As a plug-in pretext task, SP is seamlessly incorporated into MAE, MoCo v3, and DINO. (a) **MAE + SP:** To integrate SP into MAE, we share the base MAE encoder $\psi(\cdot)$. During a forward pass, the baseline MAE computes the reconstruction loss $\mathcal{L}_{\text{MAE}}$ using the masked image I. Concurrently, the SP branch takes two separate augmented views v_1 and v_2 as input. The shared encoder $\psi(\cdot)$ processes v_1 and v_2 to output the corresponding patch tokens $\mathbf{Z}_{v1}$ and $\mathbf{Z}_{v2}$. The Spatial Predictor takes these tokens to compute the spatial regression loss $\mathcal{L}_{\text{SP}} = \mathcal{L}_{pos} + \mathcal{L}_{scale}$. (b) **MoCo v3 + SP:** When incorporating SP into MoCo v3, we specifically share the **online encoder** $\psi_{\text{online}}(\cdot)$ as the feature extractor for the SP branch. While the baseline computes the MoCo loss $\mathcal{L}_{\text{MoCo}}$ using the similarities between the online and momentum projections of x_1 and x_2, the SP branch processes its own views v_1 and v_2 through the shared $\psi_{\text{online}}(\cdot)$. This design forces the online encoder to not only achieve invariance to heavy augmentations (via $\mathcal{L}_{\text{MoCo}}$) but also maintain an explicit understanding of spatial equivariance (via $\mathcal{L}_{\text{SP}}$). The gradients from both $\mathcal{L}_{\text{MoCo}}$ and $\mathcal{L}_{\text{SP}}$ are backpropagated through the online encoder. (c) **DINO + SP:** For integration with DINO, the SP branch specifically shares the **student encoder** $\psi_{\text{student}}(\cdot)$. The views v_1 and v_2 are fed into $\psi_{\text{student}}(\cdot)$ to extract the class tokens $\mathbf{z}_{v1}$ and $\mathbf{z}_{v2}$ (and/or patch tokens), which are then passed to the Spatial Predictor. Crucially, the gradients from the SP auxiliary loss $\mathcal{L}_{\text{SP}}$ only update the student encoder, while the teacher encoder remains updated solely via exponential moving average (EMA).

Table S1: **Pre-training hyperparameters for 100-epoch evaluation.** We maintain identical optimization settings between each baseline and its $+SP$ variant to ensure a fair comparison. All models are trained on ImageNet-1K with a total batch size of 256. We borrow code from sololearn da Costa et al. (2022), lightlySusmelj et al., and DINOCaron et al. (2021); Oquab et al. (2024); Siméoni et al. (2025)

Hyperparameters	MAE (ViT-B)		MoCo v3 (ViT-S)		DINO (ViT-S)	
	Baseline	+SP	Baseline	+SP	Baseline	+SP
Optimization						
Optimizer	AdamW		AdamW		AdamW	
Base Learning Rate	1.5e-4		1.5e-4		5e-4	
Weight Decay	0.05		0.1		$0.04 \rightarrow 0.4$	
Optimizer Momentum	$\beta_{1,2}$=(0.9, 0.95)		$\beta_{1,2}$=(0.9, 0.999)		$\beta_{1,2}$=(0.9, 0.999)	
Training Schedule						
Total Epochs	100		100		100	
Batch Size	256		256		256	
Learning Rate Schedule	Cosine decay					
Warmup Epochs	10		10		10	
Method-Specific						
Masking Ratio	75%	75%	N/A	N/A	N/A	N/A
EMA Momentum	N/A	N/A	$0.99 \rightarrow 1.0$		$0.996 \rightarrow 1.0$	
Temperature (τ)	N/A	N/A	0.2		$0.04 \rightarrow 0.07$	
Multi-crop Scale	N/A	N/A	N/A	N/A	(0.4, 1.0) / (0.05, 0.4)	
SP Loss Weight (λ)	–	0.1	–	0.1	–	0.1

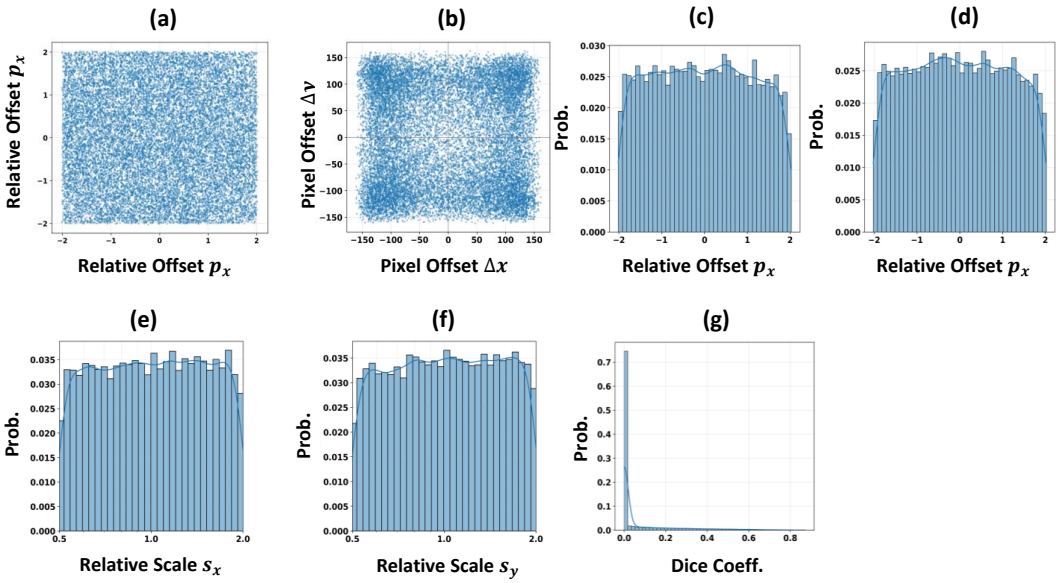

Figure S2: **Statistical validation of the spatial view sampling strategy.** Our sampling strategy is designed with three objectives: (i) generating semantically meaningful local views Caron et al. (2021); Zhou et al. (2022); Assran et al. (2023), (ii) inducing a broad and approximately uniform distribution over spatial prediction targets (relative position and scale), and (iii) avoiding degenerate pairs with excessive overlap. To verify that our rejection sampling accurately reproduces the intended distributions while respecting the image boundaries, we simulated the generation of $N = 20{,}000$ pairs of local views. **(a)** Joint distribution of relative offsets (p_x, p_y). **(b)** Scatter plot of raw pixel offsets $(\Delta x, \Delta y)$. **(c, d)** Marginal distributions of p_x and p_y. **(e, f)** Log-uniform distributions of relative scales s_x and s_y. **(g)** Distribution of pairwise view Dice overlap. The sampled relative offsets and log-scale ratios approximately follow the intended uniform distributions in (a–f), indicating that the rejection sampling procedure preserves the target distributions despite finite image boundary constraints. Furthermore, the Dice overlap distribution in (g) is concentrated near zero, suggesting that the sampler effectively suppresses highly overlapping and potentially redundant view pairs while maintaining spatial diversity.

Table S2: **Ablation on spatial loss weight λ.** We report the CIFAR-100 linear probing (LP) top-1 accuracy. The experiments are conducted using the MoCoV3 (ViT-Tiny) backbone pre-trained for 800 epochs. Setting $\lambda_p = \lambda_s = 0.1$ yields the best balance between semantic representation and spatial reasoning.

Spatial Loss Weight ($\lambda_p = \lambda_s$)	**CIFAR-100 LP Acc (%)**
$\lambda_p = \lambda_s = 0.05$	63.5
$\lambda_p = \lambda_s = 0.1$ (Ours)	**64.8**
$\lambda_p = \lambda_s = 0.5$	62.7

Table S3: **Additional in-domain classification comparisons.** We compare all methods on CIFAR-100 using a ViT-Tiny backbone and evaluate the learned representations under the same linear probing (LP) protocol. We first compare with PART Ayoughi et al. (2025), which directly uses spatial information as supervision for representation learning. In contrast, our SP is introduced as an auxiliary plug-in to MoCo v3, retaining the original MoCo v3 objective while adding a spatial prediction objective through an SP-specific decoder. We further include MoCo v3 + Extra Crops to control for the additional crops introduced by SP. This baseline uses the same two additional crops as SP, but removes the SP decoder and spatial prediction objective; instead, these crops are processed by the MoCo v3 backbone and aligned with the original crops using the standard MoCo objective. MoCo v3 + SP achieves the highest accuracy, outperforming both PART and the extra-crop control. These results show that SP can effectively enhance an existing SSL framework and that its improvement cannot be explained simply by introducing additional crops.

Method	**CIFAR-100 LP Acc (%)**
PART	61.9
MoCo v3 + Extra Crops	63.6
MoCo v3 + SP (Ours)	**64.8**

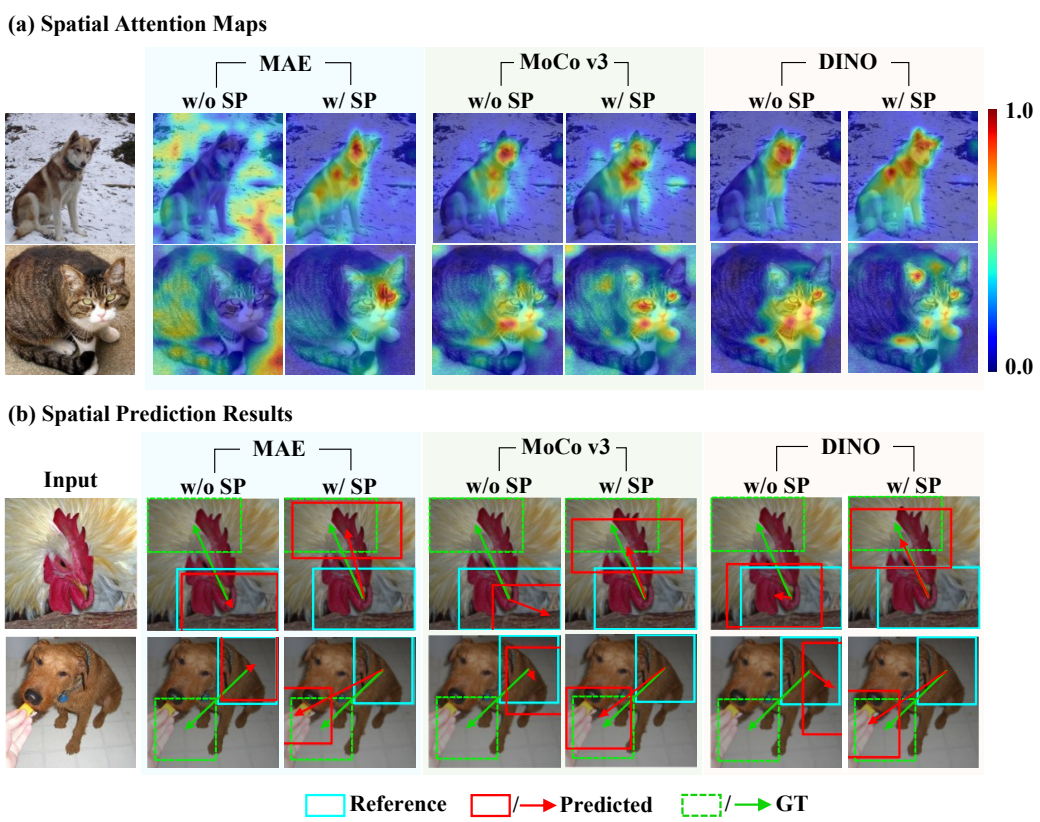

Figure S3: **Additional visualization examples for MoCo v3 with and without SP.** We provide additional qualitative examples following the same design convention as **Fig.3** in the main text.

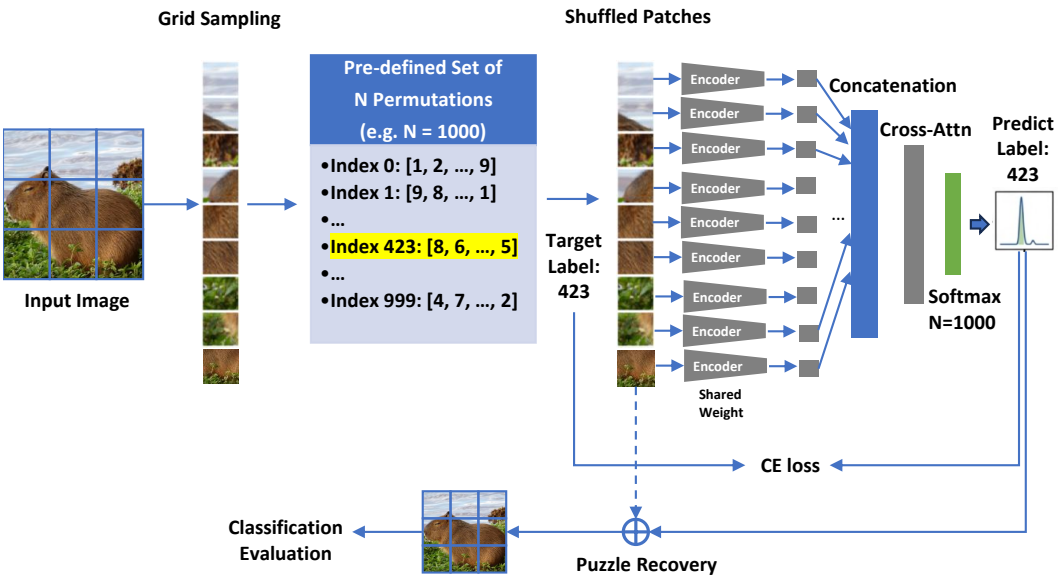

Figure S4: **Overview of the Jigsaw Understanding Evaluation.** To assess the joint modeling of spatial and semantic information, we construct a jigsaw reconstruction task following the formulation of Noroozi & Favaro (2016). Given an input image, nine local patches are sampled and shuffled according to a target index from a predefined set of $N = 1,000$ permutations. These permutations are selected to maximize pairwise Hamming distance, thereby mitigating shortcut learning. Each patch is independently processed through a shared-weight encoder to extract semantic features. In our design, a Cross-Attention mechanism is employed to enable each patch query to reason about its relative position conditioned on the global context of all other patches. The head is trained using Cross-Entropy (CE) loss to predict the permutation index. During evaluation, we perform puzzle recovery based on the predicted index and feed the reconstructed image back into the frozen pre-trained backbone with its linear probe to obtain classification performance.