

UNSUPERVISED LEARNING WITH SPATIAL EMBEDDING AND HUMAN LABELING

Anonymous authors

Paper under double-blind review

ABSTRACT

Large-scale foundation models such as CLIP and DINOv2 provide powerful pre-trained visual embeddings that enable strong zero-shot transfer and facilitate unsupervised learning. However, for specific tasks, the visual embeddings extracted from these foundation models may still lack sufficient classification separability, making it challenging to identify a reliable classifier in the embedding space. To address this, we propose an unsupervised learning approach with spatial embedding and human labeling (SEAL). SEAL first extracts spatial embeddings using a graph attention network to capture relational cues among image patches. These spatial embeddings are then fused with foundation model features via mutual distillation, producing spatially aware representations with enhanced separability. Subsequently, a lightweight linear classifier is trained in this embedding space to generate cluster assignments that reflect human labeling. Experimental analysis on 26 benchmark datasets shows that incorporating spatial embeddings significantly improves triplet accuracy, demonstrating the enhanced separability of foundation model embeddings. Extensive experiments further show that SEAL achieves outstanding clustering performance across 26 benchmark datasets and maintains excellent stability across 7 foundation model backbones. The code will be released publicly.

1 INTRODUCTION

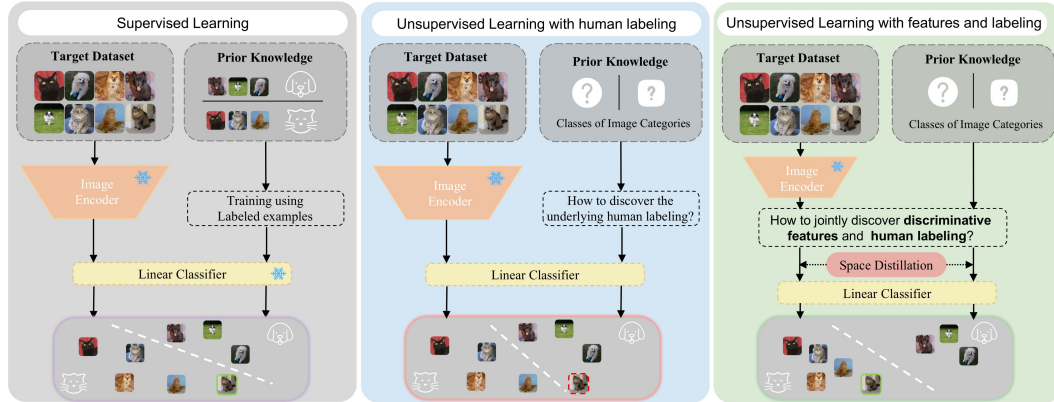
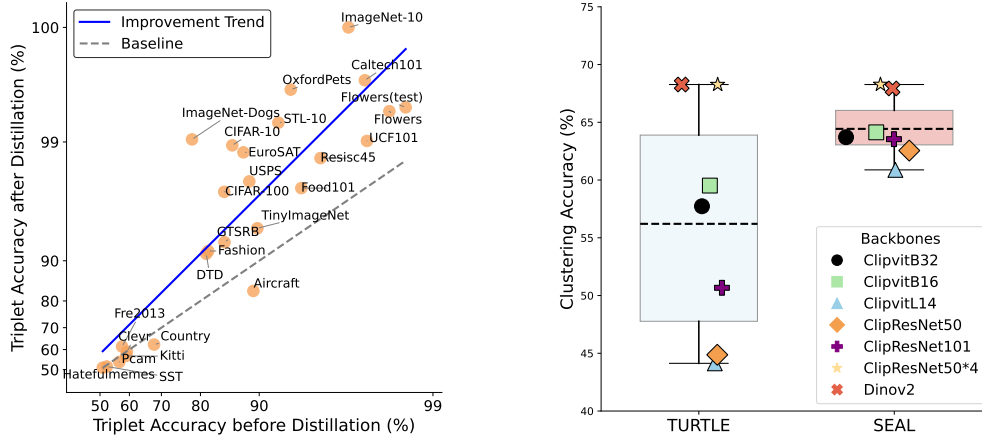


Figure 1: Comparison of Supervised and Unsupervised Learning Approaches. (a) Supervised learning trains models on labeled examples with predefined image categories. (b) Unsupervised learning infers underlying human labeling without labeled examples, relying only on the specified number of categories. (c) Unsupervised learning simultaneously discovers discriminative features and recovers human labeling, only given the number of categories.

Deep learning has revolutionized computer vision by enabling models to learn representations with rich information directly from raw images (LeCun et al., 2015). In supervised learning (illustrated on the left of Figure 1), labeled datasets provide the critical training signal (Krizhevsky et al., 2012;



(a) Triplet accuracy comparison between different datasets with and without spatial structure using ClipVIT-L/14. The blue line shows the overall improvement trend, and the gray dashed line indicates similar triplet accuracy before and after distillation.

(b) Boxplots depict the distribution of the average clustering accuracies on 26 datasets for TURTLE (blue) and SEAL (red) across 7 backbones. Individual points indicate the average accuracy achieved by each backbone.

Figure 2: (a) Triplet accuracy before and after distillation. (b) Clustering performance across backbones.

He et al., 2016; Ren et al., 2015), and the learned representations serve as the foundation for achieving excellent performance on core tasks such as image classification, object detection, and semantic segmentation. In parallel, unsupervised representation learning aims to uncover meaningful information from raw images without relying on human annotations. Early representative methods in this field, such as Deep Embedded Clustering (DEC) (Xie et al., 2016) and Contrastive Clustering (CC) (Li et al., 2021), have shown that leveraging the intrinsic properties of the data enables models to learn transferable embeddings that generalize effectively across downstream tasks.

With the accumulation of data across multiple domains, numerous large-scale foundation models have emerged, including CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2023), SWAG (Singh et al., 2022), and CoCa (Yu et al., 2022). Their pre-trained visual embeddings have evolved into powerful priors for a wide range of downstream tasks, exhibiting strong zero-shot transfer capabilities by aligning input data with human-provided prompts. However, reliance on such external supervision limits their applicability in fully unsupervised scenarios. To address this limitation, recent methods such as HUME (Gadetsky & Brbic, 2023) and TURTLE (Gadetsky et al., 2024) reinterpret clustering as the task of discovering human-consistent labelings from pre-trained embeddings. The core insight behind the methods is that human-defined categories tend to be linearly separable in the representation space. This property allows for searching through potential labelings and evaluating them based on the performance of linear classifiers. Using this principle, these methods aim to recover semantically meaningful data partitions without relying on human-provided prompts, thereby advancing toward fully unsupervised utilization of foundation models.

However, for specific tasks, the pre-trained embeddings, despite containing rich semantic information, may still lack sufficient linear separability, making it challenging to identify a reliable linear classifier in the feature space.

To address this limitation, we propose a method to enhance the linear separability of visual embeddings extracted from foundation models. Our approach starts with a Graph Attention Network (GAT) encoder, which extracts spatial embeddings by capturing relational cues between image patches. These spatial embeddings are then fused with the visual embeddings of the foundation via mutual distillation to generate spatially aware embeddings that better retain the structural information of images. To quantify the improvement of the spatially aware embeddings in separability, we adopt a triplet accuracy, which evaluates the relative closeness of samples from the same class versus those from different classes, thereby reflecting the linear separability of features (see Appendix D for details). As illustrated in Figure 2a, integrating spatial structure into foundation model embeddings

yields significantly higher triplet accuracy, showing that the linear separability of features is effectively enhanced. We then discover actual clustering assignments for the data based on the spatially aware embeddings.

Building on the above, we introduce an unsupervised learning method with spatial embedding and human labeling (SEAL), which integrates spatial structures to enhance feature separability and recover the underlying human labeling simultaneously. Following the process detailed earlier (i.e., spatial embedding extraction via GAT and mutual distillation with visual embeddings from foundation models), inspired by (Gadetsky et al., 2024), a lightweight linear classifier is trained on this spatially aware embedding space to output cluster assignments. We evaluated the performance of SEAL on 26 benchmark datasets and 7 foundation model backbones (including CLIP ResNets, CLIP Vision Transformers, and DINOv2). As illustrated in Figure 2b, SEAL outperforms the TURTLE baseline. In particular, it achieves more stable performance across diverse foundation model backbones. Detailed results for each dataset are provided in Appendix C. In addition, although spatial embedding consumes extra time, the clearer class structure improves the efficiency of recovering human labeling, and the efficiency of SEAL is still at a comparable level. These results confirm that explicitly modeling spatial structures provides dual benefits: it enhances the linear separability of features (as validated by triplet accuracy) and improves the effectiveness and stability of clustering, thereby advancing the state-of-the-art in unsupervised learning.

The main contributions are as follows:

- We propose a spatially aware approach that generates spatial features and undergoes mutual distillation with embeddings from foundation models, thereby capturing spatial structure relationships and enhancing separability. The improvement in separability is quantitatively demonstrated across 26 datasets and 7 backbones using triplet accuracy.
- We propose an unsupervised learning method with spatial embeddings and human labeling (SEAL), which recovers human labeling based on spatially aware embeddings. Extensive experiments demonstrate that SEAL achieves outstanding clustering performance on 26 benchmark datasets and exhibits excellent stability across 7 foundation model backbones.

2 RELATED WORK

2.1 REPRESENTATION LEARNING FOR IMAGE CLUSTERING

Classical methods such as K-means or spectral clustering (Krishna & Murty, 1999; Ng et al., 2001) rely on low-level embeddings and often fail when applied to large-scale datasets with complex semantics.

With the advent of deep learning, image clustering performance has significantly improved by leveraging neural networks. These methods rely on internal supervision signals to guide the learning of clustering-friendly representations. They typically assume that samples from the same category are naturally closer in the embedding space and optimize the representations accordingly. Early works like Deep Embedded Clustering (DEC) (Xie et al., 2016) refine cluster assignments by minimizing the KL divergence between current predictions and a sharpened target distribution. Prediction-consistency based approaches such as Invariant Information Clustering (IIC) (Ji et al., 2019), SCAN (Van Gansbeke et al., 2020), and Contrastive Clustering (CC) (Li et al., 2021) leverage agreement under strong data augmentations to facilitate unsupervised clustering. Alternatively, iterative pseudo-labeling methods, including DeepCluster (Caron et al., 2018) and SPICE (Niu et al., 2022), repeatedly assign cluster labels to supervise representation learning. Graph-based extensions like GATCluster (Niu et al., 2020) introduce graph attention mechanisms to model neighborhood dependencies and improve clustering quality.

2.2 UNSUPERVISED LEARNING WITH FOUNDATION MODELS

While deep clustering methods learn task-specific embeddings, recent works leverage large-scale pretrained models to provide rich visual representations, with CLIP and CoCa capturing high-level semantic concepts through image-language alignment, and DINOv2 and SWAG excelling at preserving visual details. While these models demonstrate strong zero-shot transfer, their reliance on external supervision limits applicability in fully unsupervised settings.

To address this, recent works such as HUME (Gadetsky & Brbic, 2023) and TURTLE (Gadetsky et al., 2024) reframe clustering as the task of discovering human labeling directly from pretrained embeddings. HUME introduced the idea of searching for labelings that are linearly separable across multiple representation spaces, using self-supervised features for the task encoder and a large pre-trained model as a regularizer. However, it still required task-specific self-supervised pretraining on the target dataset. TURTLE advances this paradigm by entirely removing the need for dataset-specific representation learning, instead leveraging fixed embeddings from one or two foundation models to define the task encoder and evaluate generalization error. By maximizing the margin of linear classifiers across these spaces, TURTLE effectively identifies labelings that closely align with human labeling, often matching or surpassing zero-shot transfer methods’ performance without any human labeling.

These developments highlight a broader shift in unsupervised learning: moving from learning embeddings from scratch to leveraging pretrained foundation models to recover semantically meaningful data partitions. This trend improves scalability and efficiency, and narrows the gap between fully unsupervised and supervised learning. Nevertheless, the lack of sufficient class separability remains a persistent challenge in unsupervised learning.

3 THE PROPOSED METHOD

3.1 PROBLEM STATEMENT AND OBJECTIVE

In this paper, we consider the following unsupervised learning problem: given an unlabeled training dataset $\mathcal{D} = \{x_n\}_{n=1}^N$ with a known number of categories K , the goal is to learn a model $f : \mathcal{D} \rightarrow \{1, \dots, K\}$ that assigns a pseudo-label $y_i = f(x_i)$ to each sample x_i , such that samples belonging to the same class exhibit homogeneity.

We address this problem through spatial embedding and human labeling. Each image x is represented by a pre-trained visual embedding $\phi(x) \in \mathbb{R}^d$. Our approach has two main objectives: (i) to enhance the linear separability of the visual embeddings, and (ii) to infer a human-consistent labeling distribution $\tau_\theta(x) \in \Delta^{K-1}$, which assigns semantically meaningful category probabilities to each sample, where Δ^{K-1} denotes the $(K-1)$ -dimensional probability simplex. Formally, we leverage additional spatial embeddings $g(x) \in \mathbb{R}^d$. By integrating these spatial embeddings with the visual embeddings $\phi(x)$, we aim to produce spatially aware embeddings $\hat{\phi}(x)$ that improve linear separability and make clustering categories more distinguishable in the feature space. Subsequently, we learn a mapping $\hat{\phi}(x) \in \mathbb{R}^d$ such that a linear classifier $f(\hat{\phi}(x)) = w^\top \hat{\phi}(x)$ assigns samples to the K clusters according to the inferred distribution $\tau_\theta(x)$. An overview of the proposed framework is illustrated in Figure 3.

Guided by the above solution, we propose a learnable distillation function $\mathcal{R}$ that integrates $\phi(x)$ and $g(x)$ to obtain spatially-aware embeddings $\hat{\phi}(x) = \mathcal{R}(\phi(x), g(x))$. A classifier $f(x) = w^\top \hat{\phi}(x)$ is then trained on this embedding space. The overall optimization problem is formulated as:

$$\min_{\theta} \mathcal{L}(\theta) = \sum_{x \in \mathcal{D}} \mathcal{L}_{ce}(f(x), \tau_\theta(x)) - \beta \mathcal{L}_{ent}(\tau_\theta(x)), \quad (1)$$

$$\text{s.t.} \quad \left\{ \begin{array}{l} \hat{\phi}(x) = \mathcal{R}(\phi(x), g(x)), \\ \mathcal{R} = \arg \min_{\mathcal{R}'} \mathcal{L}_{\text{distill}}(\phi(x), g(x); \mathcal{R}') \\ f(x) = w^\top \hat{\phi}(x), \\ w = \arg \min_{w'} \mathcal{L}_{ce}(w'^\top \hat{\phi}(x), \tau_\theta(x); w') \end{array} \right\} \begin{array}{l} \text{spatially aware embedding} \\ \text{human labeling} \end{array} \quad (2)$$

where β is a weighting factor for the entropy loss of the latent label distribution, and $\mathcal{L}_{\text{ent}}^{\tau_\theta}$ denotes the entropy loss of $\tau_\theta(x)$. The spatial embedding $g(x)$ is obtained as described in Section 3.2 and the learning of $\mathcal{R}$ is detailed in Section 3.3. The human labeling $f(x)$ on spatially aware embeddings is introduced in Section 3.4. The entire framework is trained by jointly optimizing θ , $\mathcal{R}$, and w .

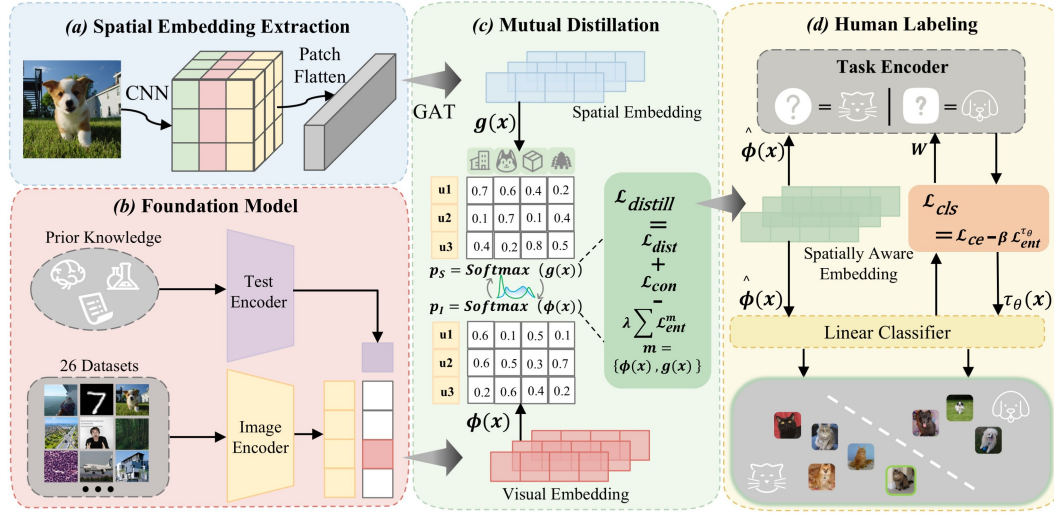


Figure 3: Overview of the proposed SEAL framework.

3.2 SPATIAL EMBEDDING EXTRACTION

Capturing spatial structure is crucial for distinguishing task-specific categories. Inspired by the structural representation idea in (Qian et al., 2015), we model spatial dependencies among image patches using GAT (Veličković et al., 2017). Since raw image dataset $\mathcal{D}$ cannot be directly input to GAT, We first feed $\mathcal{D}$ into a pretrained ResNet-50 (Radford et al., 2021) to extract high-level convolutional features, which are then used as inputs for the GAT, including node embeddings $X = [x_1, \dots, x_N]^\top \in \mathbb{R}^{N \times d}$ and edge information $E = \{(i, j) \mid i \neq j, i, j \in \{1, \dots, N\}\}$.

Then, X and E are input to the GAT, which updates each node by attending to its neighbors:

$$x'_i = \sum_{j \in \mathcal{N}(i)} \alpha_{ij} W x_j, \quad \alpha_{ij} = \frac{\exp \left(\text{LeakyReLU} \left(a^\top [W x_i \parallel W x_j] \right) \right)}{\sum_{k \in \mathcal{N}(i)} \exp \left(\text{LeakyReLU} \left(a^\top [W x_i \parallel W x_k] \right) \right)}, \quad (3)$$

where W is a learnable weight matrix, $\mathcal{N}(i)$ denotes the neighbors of node i , a is a learnable vector, and $\parallel$ denotes vector concatenation.

After refining the node embeddings, we apply global average pooling to $X' = [x'_1, \dots, x'_N]^\top \in \mathbb{R}^{N \times d}$ in order to generate the spatial embedding:

$$g(x) = \frac{1}{N} \sum_{i=1}^N x'_i \in \mathbb{R}^d. \quad (4)$$

In the following section, we fuse the spatial embedding $g(x)$ derived above with the visual embedding $\phi(x)$ to obtain spatially aware embeddings.

3.3 MUTUAL DISTILLATION

Knowledge distillation (KD) (Hinton et al., 2015) has been shown to enhance embedding consistency through cross-model alignment (Radford et al., 2021). Motivated by this, we introduce a mutual distillation framework to obtain spatially aware embeddings by jointly leveraging both visual and spatial modalities.

Given an input image x , we first extract visual features $\phi(x)$ using the foundation model and obtain spatial features $g(x)$ as described in the previous section. The logits for the visual and spatial modalities are then computed via their respective clustering heads:

$$\text{logit.I} = h_I(\phi(x)), \quad \text{logit.S} = h_S(g(x)), \quad (5)$$

where h_I and h_S denote MLP projection heads for visual and spatial modalities, respectively. The spatially aware embeddings are subsequently obtained by integrating the two modalities through a learnable distillation function (corresponding to the spatially aware embedding component in Eq. 2):

$$\hat{\phi}(x) = \mathcal{R}(\phi(x), g(x)), \quad \mathcal{R} = \arg \min_{\mathcal{R}'} \mathcal{L}_{\text{distill}}(\phi(x), g(x); \mathcal{R}'). \quad (6)$$

The optimization objective $\mathcal{L}_{\text{distill}}$ combines three terms, which are distillation loss, consistency loss, and entropy loss. The mutual distillation loss aligns the cluster distributions between visual and spatial modalities:

$$\mathcal{L}_{\text{dist}} = \mathcal{L}_{S \rightarrow I} + \mathcal{L}_{I \rightarrow S}, \quad (7)$$

$$\mathcal{L}_{S \rightarrow I} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{\text{sim}(\text{logit}_I, \text{logit}_S)/T}}{\sum_{j=1}^N e^{\text{sim}(\text{logit}_I, \text{logit}_S_j)/T}}, \quad (8)$$

$$\mathcal{L}_{I \rightarrow S} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{\text{sim}(\text{logit}_S, \text{logit}_I)/T}}{\sum_{j=1}^N e^{\text{sim}(\text{logit}_S, \text{logit}_I_j)/T}}, \quad (9)$$

where T is a temperature parameter. The consistency loss enforces agreement between modalities:

$$\mathcal{L}_{\text{con}} = -\frac{1}{N} \sum_{i=1}^N \log \left(\text{logit}_I^\top, \text{logit}_S \right). \quad (10)$$

The entropy loss prevents degenerate solutions:

$$\mathcal{L}_{\text{ent}}^m = -\frac{1}{N} \sum_{i=1}^N \frac{1}{K} \sum_{j=1}^K (\text{logit}_{i,j}^m) \log(\text{logit}_{i,j}^m), \quad m \in \{\phi(x), g(x)\}. \quad (11)$$

The overall training objective is defined as:

$$\mathcal{L}_{\text{distill}} = \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{con}} - \lambda \sum_m \mathcal{L}_{\text{ent}}^m. \quad (12)$$

where the entropy regularization weight is fixed to $\lambda = 5$.

After the mutual distillation, we obtain the spatially aware embeddings from the distilled image clustering head. These embeddings preserve the spatial structural relationships of the images. In the next subsection, we leverage the spatially aware embedding space to discover the underlying human labeling, thereby generating the cluster assignment.

3.4 HUMAN LABELING

After obtaining the spatially aware embeddings $\hat{\phi}(x)$ via mutual distillation, we freeze them and adopt the bi-level optimization protocol of TURTLE (Gadetsky et al., 2024) to train a final linear classifier.

Specifically, given the pseudo-label distribution $\tau_\theta(x) \in \Delta^{K-1}$ produced by the task encoder at the t -th outer iteration, the linear classifier is updated in the inner loop on the current mini-batch $\mathcal{B}^{(t)}$ via 10 steps of SGD:

$$w = \arg \min_{w'} \mathcal{L}_{\text{ce}}(w'^\top \hat{\phi}(x), \tau_\theta(x); w'), \quad f(x) = w^{(t)\top} \hat{\phi}(x) \in \mathbb{R}^K. \quad (13)$$

In the outer loop, the task encoder parameters θ are updated by minimizing $\mathcal{L}_{\text{ce}}$ with an additional entropy regularization term:

$$\min_{\theta} \mathcal{L}(\theta) = \sum_{x \in \mathcal{D}} \mathcal{L}_{\text{ce}}(f(x), \tau_\theta(x)) - \beta \mathcal{L}_{\text{ent}}(\tau_\theta(x)), \quad (14)$$

where the entropy regularization weight is fixed to $\beta = 10$.

Following TURTLE, the linear classifier can operate on spatially-aware embeddings generated from either one or two backbones, which we denote as 1-space and 2-space, respectively. For the 2-space case, the per-space $\tau_\theta(x)$ are first computed and then averaged to obtain the final label assignments.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets and evaluation metric. The experimental analysis is conducted on 26 benchmark datasets spanning object recognition, scene understanding, fine-grained classification, and remote sensing. The detailed description of each dataset is provided in Appendix A.1. Train and test splits follow the original dataset configurations. To evaluate the performance, we adopted Accuracy (ACC) (Xie et al., 2016), Normalized Mutual Information (NMI) (Estévez et al., 2009) and Adjusted Rand Index (ARI) (Steinley, 2004). All experiments are repeated ten times with different random seeds, and the average indices values are reported.

Baselines. We employ a range of state-of-the-art clustering methods as reference, including K-Means (Krishna & Murty, 1999), DEC (Xie et al., 2016), DAC (Chang et al., 2017), DSEC (Chang et al., 2018), DCCM (Wu et al., 2019), MICE (Tsai et al., 2020), GATCluster (Niu et al., 2020), PLCA (Huang et al., 2020), IDFD (Tao et al., 2021), CC (Li et al., 2021), C3 (Sadeghi et al., 2022), TCL (Li et al., 2022), CONCUR (Deshmukh et al., 2022), SPICE (Niu et al., 2022), TAC (Li et al., 2023), DPAC (Yan et al., 2024), and TURTLE (Gadetsky et al., 2024).

Foundation Representations. We employ CLIP (Radford et al., 2021) representations across different architectures and model sizes, including three ResNets (ResNet50, ResNet101, ResNet50x4), three Vision Transformers (ViT-B/32, ViT-B/16, and ViT-L/14), and DINOv2 ViT-G/14 (Oquab et al., 2023). For clarity, we refer to SEAL using a representation space generated from a single backbone as 1-Space, and SEAL combining two representation spaces as 2-Space. The specific backbones employed in each experiment are indicated in the corresponding sections. Additional details on the models and the procedure for preparing representations are provided in Appendix A.2.

Implementation Details. We implement Mutual Distillation with two MLP-based cluster heads. For the features of one modality, we select the top-50 nearest neighbors from the other modality’s features to calculate loss. The cluster heads are trained with Adam ($\text{lr} = 0.001, \gamma = (0.9, 0.99)$) for 20 epochs (batch size 512) and temperature $t = 1.0$. Input embeddings have backbone-dependent dimensions, and the final spatially aware embeddings retain the same dimensionality as the input. We implement linear classifier using a bi-level optimization scheme. The outer loop runs for $T = 6000$ iterations with Adam ($\text{outer_lr} = 0.001, \gamma = (0.9, 0.999)$) and batch size 10,000, updating the task encoder. The inner loop performs $M = 10$ steps per outer iteration with Adam ($\text{inner_lr} = 0.001, \gamma = (0.9, 0.999)$), updating the weight parameters w of the linear classifier.

4.2 EXPERIMENTAL RESULTS

4.2.1 COMPARATIVE ANALYSIS RESULTS

For clarity, in this comparison, TURTLE (1-space) and SEAL (1-space) both use features extracted from the DINOv2 backbone, whereas TURTLE (2-space) and SEAL (2-space) leverage dual-space features from CLIP ViT-L/14 and DINOv2 backbones.

The results on 4 widely used benchmark datasets are reported in Table 1. SEAL achieves state-of-the-art performance on ImageNet-10, ImageNet-Dogs, and STL-10, with nearly perfect clustering results on ImageNet-10 and STL-10 datasets. On CIFAR-10, SEAL performs competitively and remains close to TURTLE (Gadetsky et al., 2024), leading to the best overall average ACC performance across all methods. In this comparison, the performance of SEAL (1-space) is superior to that of TURTLE (1-space). More comparison experiments between SEAL, K-Means, and TURTLE can be found in Appendices E and F. These results illustrate the effectiveness of SEAL.

4.2.2 TIME ANALYSIS

We evaluate the efficiency of our method versus TURTLE on four datasets. Figure 4 shows total runtimes for increasing spatial structure and recovering human labeling using three backbones: CLIP ViT-L/14, DINOv2, and CLIP ViT-L/14 + DINOv2. Although increasing spatial structure adds some overhead, it clarifies class structure and accelerates human labeling recovery, making SEAL’s overall efficiency comparable to TURTLE; in some cases, the speedup even offsets the extra time.

Table 1: Clustering performance of different approaches evaluated by ACC%, NMI%, and ARI%. “–” indicates unavailable results. **Bold** numbers indicate the best performance across all methods. Underlined numbers indicate the best performance when comparing between TURTLE (1-space) vs SEAL (1-space) and TURTLE (2-space) vs SEAL (2-space), respectively.

Method	ImageNet-10			ImageNet-Dogs			STL-10			CIFAR-10			Avg.
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	
K-Means (Krishna & Murty, 1999)	24.1	11.9	5.7	10.5	5.5	2.0	19.2	12.5	6.1	22.9	8.7	4.9	19.2
DEC (Xie et al., 2016)	38.1	28.2	20.3	19.5	12.2	7.9	35.9	27.6	18.6	30.1	25.0	16.1	30.9
DAC (Chang et al., 2017)	52.7	39.4	30.2	27.5	21.9	11.1	47.0	36.6	25.7	52.2	40.0	30.1	44.9
DSEC (Chang et al., 2018)	67.4	58.3	52.2	26.4	23.6	12.4	48.2	40.3	28.6	47.8	43.8	34.0	47.5
DCCM (Wu et al., 2019)	71.0	60.8	55.5	38.3	32.1	18.2	48.2	37.6	26.2	62.3	49.6	40.8	54.9
MiCE (Tsai et al., 2020)	–	–	–	43.9	42.3	28.6	75.2	63.5	57.5	83.5	73.7	69.8	67.5
GATCluster (Niu et al., 2020)	76.2	60.9	57.2	33.3	32.2	20.0	58.3	44.6	36.3	61.0	47.5	40.2	57.2
PICA (Huang et al., 2020)	87.0	80.2	76.1	35.2	35.2	20.1	71.3	61.1	53.1	69.6	59.1	51.2	65.8
IDFD (Tao et al., 2021)	95.4	89.8	90.1	59.1	54.6	41.3	75.6	64.3	57.5	81.5	71.1	66.3	77.9
CC (Li et al., 2021)	89.3	85.9	82.2	42.9	44.5	27.4	85.0	76.4	72.6	79.0	70.5	63.7	74.1
C3 (Sadeghi et al., 2022)	94.2	90.5	86.1	43.4	44.8	28.0	–	–	–	83.8	74.8	70.7	73.8
TCL (Li et al., 2022)	89.5	87.5	83.7	64.4	62.3	51.6	86.8	79.9	75.7	88.7	81.9	78.0	82.4
ConCUR (Deshmukh et al., 2022)	95.8	90.7	90.9	69.5	63.0	53.1	74.9	63.6	56.6	84.6	76.2	71.5	81.2
SPICE (Niu et al., 2022)	96.9	92.7	93.3	67.5	62.7	52.6	92.9	86.0	86.5	91.8	85.0	83.6	87.3
TAC (Li et al., 2023)	99.4	98.5	98.8	84.4	77.4	72.0	98.3	95.7	96.3	92.2	83.7	83.6	93.6
DPAC (Yan et al., 2024)	97.0	92.5	93.5	72.6	66.7	59.8	93.4	86.3	86.1	93.4	87.0	86.6	89.1
TURTLE(1-space)(Gadetsky et al., 2024)	89.1	88.5	83.0	87.0	84.8	78.7	56.2	58.0	41.3	90.8	88.4	83.8	80.8
SEAL(1-space)	<u>99.7</u>	<u>99.1</u>	<u>99.3</u>	<u>97.8</u>	<u>95.1</u>	<u>95.4</u>	<u>87.1</u>	<u>86.6</u>	<u>78.7</u>	<u>98.6</u>	<u>96.1</u>	<u>96.8</u>	<u>95.8</u>
TURTLE(2-space)(Gadetsky et al., 2024)	99.8	99.3	99.5	91.5	88.9	85.5	99.8	99.4	99.5	99.5	98.3	98.8	97.6
SEAL(2-space)	<u>99.9</u>	<u>99.6</u>	<u>99.7</u>	<u>97.9</u>	<u>95.3</u>	<u>95.5</u>	<u>99.9</u>	<u>99.8</u>	<u>99.9</u>	98.7	96.5	97.2	<u>99.1</u>

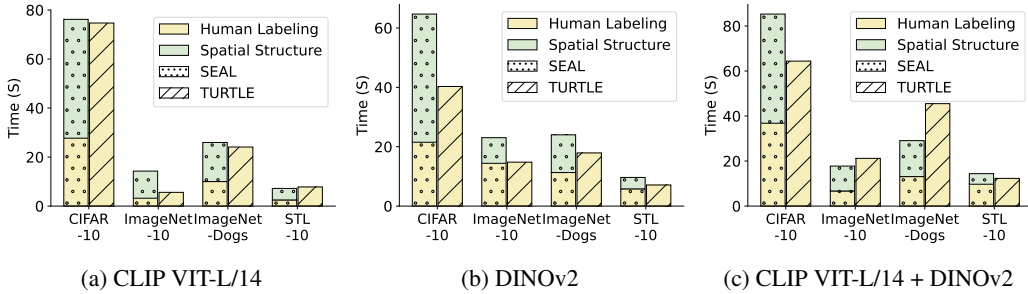


Figure 4: Time consumption of SEAL and TURTLE across 4 datasets under 3 backbones.

4.3 EFFECTIVENESS OF SPATIALLY AWARE EMBEDDINGS

To evaluate the contribution of spatially aware embeddings to representation learning, we conduct an ablation study using K-Means clustering. Specifically, we compare two settings: (1) directly concatenating the original CLIP VIT-L/14 and DINOv2 embeddings (baseline), and (2) first applying the mutual distillation to both embeddings before concatenation. We employ the standard K-Means algorithm, with the number of clusters set equal to the ground-truth class count for each dataset. To reduce randomness, K-Means is repeated 30 times with different centroid initializations, using a fixed random seed to ensure reproducibility. Figure 5 presents the accuracy improvements across 26 benchmark datasets. As shown, spatially aware embeddings substantially improve clustering performance on most datasets. Detailed results for each dataset are provided in Table 7 in Appendix F.

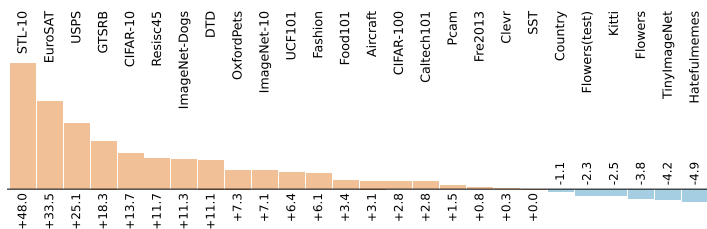


Figure 5: K-Means accuracy improvement (%) from concatenating embeddings of VIT-L/14 and DINOv2 after applying a spatially-aware transformation to each, compared to directly concatenating their original embeddings, across 26 benchmark datasets.

4.3.1 EFFECT OF MUTUAL DISTILLATION

We first evaluate the effect of Mutual Distillation (MD) on 1-space representation. Figure 6 reports the average ACC of SEAL (with MD) compared to the baseline TURTLE (without MD) across 7 representations, averaged over 26 datasets. As shown in Figure 6, SEAL consistently outperforms TURTLE in all representations. The improvement is particularly pronounced for convolutional backbones: CLIP ResNet50, CLIP ResNet101, and CLIP ResNet50*4 achieve gains of +16.8%, +17.7%, and +12.9%, respectively. Transformer-based representations (CLIP ViT-B/32, CLIP ViT-L/14) exhibit more moderate gains of +6.0%, +4.6%, and +0.01%, while DINOv2 shows a +3.8% improvement. These results demonstrate that Mutual Distillation enhances embedding representations. Detailed results for each dataset are provided in Appendix B.1, which are consistent with the trends in Figure 6. Furthermore, we evaluated the average ACC performance of 2-space representations, which are provided in the Appendix B.2. The results show the enhancement of the embedding by distillation.

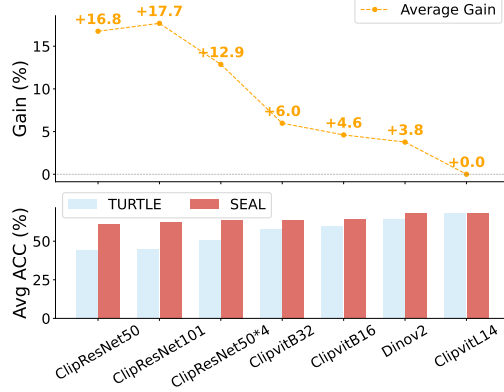


Figure 6: Effect of Mutual Distillation on 1-space representations across 26 data datasets. SEAL leverages Mutual Distillation, while TURTLE serves as the counterpart without it.

4.3.2 PARAMETER ANALYSIS

We conduct a parameter analysis of SEAL under CLIP ViT-L/14 on two representative datasets (ImageNet-10 and CIFAR-10). We vary the entropy weight ($\{4.0, 4.5, 5.0, 5.5, 6.0, 6.5\}$) and the consistency weight ($\{0, 0.5, 1.0, 1.5, 2.0, 2.5\}$), while keeping all other hyperparameters fixed. As shown in Figure 7, ACC remains stable in a wide range of parameter values. Moderate entropy regularization, combined with a consistency weight of around 1.0, generally yields good performance, demonstrating that SEAL is robust to hyperparameter variations and does not require fine-tuning.

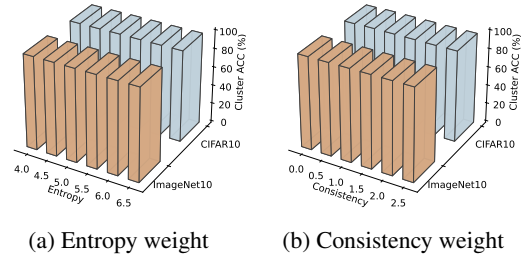


Figure 7: 3D bar plots of clustering accuracy under different entropy and consistency weights.

5 CONCLUSION

In this work, we addressed the challenge of limited linear separability in foundation model embeddings for unsupervised learning. We proposed SEAL, a framework that enhances feature separability by extracting spatial embeddings with a Graph Attention Network and integrating them with foundation model features through mutual distillation. Within this spatially aware embedding space, a linear classifier is trained to generate cluster assignments that better align with human labeling. The spatially aware embeddings improve both the structural fidelity and discriminability of representations, as evidenced by significantly higher triplet accuracy on 26 benchmark datasets. Extensive experiments on 26 datasets and 7 backbones further demonstrate that SEAL consistently achieves superior clustering performance and exhibits remarkable stability compared to strong baselines such as TURTLE. Although SEAL introduces additional computation, it improves recovering labeling efficiency by revealing clearer class structures. These results highlight the importance of incorporating spatial structures into foundation model embeddings to advance unsupervised visual representation learning. In future work, we plan to extend SEAL to multi-modal and large-scale streaming settings, further improving efficiency and adaptability in real-world applications.

REFERENCES

- Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part VI 13*, pp. 446–461. Springer, 2014.
- 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.
- Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, pp. 5879–5887, 2017.
- Jianlong Chang, Gaofeng Meng, Lingfeng Wang, Shiming Xiang, and Chunhong Pan. Deep self-evolution clustering. *IEEE transactions on pattern analysis and machine intelligence*, 42(4): 809–823, 2018.
- Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.
- Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.
- Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.
- Aniket Anand Deshmukh, Jayanth Reddy Regatti, Eren Manavoglu, and Urun Dogan. Representation learning for clustering via building consensus. *Machine Learning*, 111(12):4601–4638, 2022.
- Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.
- Pablo A Estévez, Michel Tesmer, Claudio A Perez, and Jacek M Zurada. Normalized mutual information feature selection. *IEEE Transactions on neural networks*, 20(2):189–201, 2009.
- Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.
- Artyom Gadetsky and Maria Brbic. The pursuit of human labeling: a new perspective on unsupervised learning. *Advances in Neural Information Processing Systems*, 36:60527–60546, 2023.
- Artyom Gadetsky, Yulun Jiang, and Maria Brbić. Let go of your labels with unsupervised transfer. In *International Conference on Machine Learning*, 2024.
- Andreas Geiger, Philip Lenz, and Raquel Urtasun. 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.
- Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pp. 117–124. Springer, 2013.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- Jiabo Huang, Shaogang Gong, and Xiatian Zhu. Deep semantic clustering by partition confidence maximisation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8849–8858, 2020.
- Xu Ji, Joao F. Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.
- Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ring-shia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.
- K Krishna and M Narasimha Murty. Genetic k-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(3):433–439, 1999.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 8547–8555, 2021.
- Yunfan Li, Mouxing Yang, Dezhong Peng, Taihao Li, Jiantao Huang, and Xi Peng. Twin contrastive learning for online clustering. *International Journal of Computer Vision*, 130(9):2205–2221, 2022.
- Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Jianping Fan, and Xi Peng. Image clustering with external guidance. *arXiv preprint arXiv:2310.11989*, 2023.
- Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.
- Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.
- 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.
- Chuang Niu, Jun Zhang, Ge Wang, and Jimin Liang. Gatcluster: Self-supervised gaussian-attention network for image clustering. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pp. 735–751. Springer, 2020.
- Chuang Niu, Hongming Shan, and Ge Wang. Spice: Semantic pseudo-labeling for image clustering. *IEEE Transactions on Image Processing*, 31:7264–7278, 2022.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

- Yuhua Qian, Feijiang Li, Jiye Liang, Bing Liu, and Chuangyin Dang. Space structure and clustering of categorical data. *IEEE transactions on neural networks and learning systems*, 27(10):2047–2059, 2015.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.
- Mohammadreza Sadeghi, Hadi Hojjati, and Narges Armanfard. C3: Cross-instance guided contrastive clustering. *arXiv preprint arXiv:2211.07136*, 2022.
- Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8503–8512, 2018.
- Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens Van Der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 804–814, 2022.
- Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- Johannes Stalkamp, Marc Schlipf, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012.
- Douglas Steinley. Properties of the hubert-arable adjusted rand index. *Psychological methods*, 9(3): 386, 2004.
- Yaling Tao, Kentaro Takagi, and Kouta Nakata. Clustering-friendly representation learning via instance discrimination and feature decorrelation. *arXiv preprint arXiv:2106.00131*, 2021.
- Tsung Wei Tsai, Chongxuan Li, and Jun Zhu. Mice: Mixture of contrastive experts for unsupervised image clustering. In *International conference on learning representations*, 2020.
- Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European conference on computer vision*, pp. 268–285. Springer, 2020.
- Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 210–218. Springer, 2018.
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
- Jianlong Wu, Keyu Long, Fei Wang, Chen Qian, Cheng Li, Zhouchen Lin, and Hongbin Zha. Deep comprehensive correlation mining for image clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8150–8159, 2019.
- Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
- Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pp. 478–487. PMLR, 2016.
- Yuxuan Yan, Na Lu, and Ruofan Yan. Deep online probability aggregation clustering. In *European Conference on Computer Vision*, pp. 37–54. Springer, 2024.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

APPENDIX

This appendix provides additional details for the ICLR 2026 submission, titled “Unsupervised learning with spatial embedding and human labeling”. The appendix is organized as follows:

- A Experimental details
 - A.1 Datasets
 - A.2 Foundation representations
- B Effect of mutual distillation
 - B.1 Results of 2-space Backbones on 26 datasets
 - B.2 Results of 1-space Backbone on 26 datasets
- C Clustering stability across 7 Backbones on 26 datasets
- D Separability improvement of spatial embedding
- E Performance comparison between K-Means 1-space, TURTLE 1-space, and SEAL 1-space
- F Performance comparison between K-Means 2-space, TURTLE 2-space, and SEAL 2-space

A EXPERIMENTAL DETAILS

A.1 DATASETS

We evaluate our framework on 26 benchmark datasets, covering a wide range of vision tasks. These include general object classification datasets CIFAR-10 (DeVries & Taylor, 2017), CIFAR-100 (DeVries & Taylor, 2017), STL-10 (Coates et al., 2011), TinyImageNet (Le & Yang, 2015), ImageNet-10 (Chang et al., 2017), and Caltech101 (Fei-Fei et al., 2004); fine-grained object classification datasets Food101 (Bossard et al., 2014), Flowers (Nilsback & Zisserman, 2008), Flowers(Test) (Nilsback & Zisserman, 2008), FGVC Aircraft (Maji et al., 2013), ImageNet-Dogs (Chang et al., 2017), and OxfordPets (Parkhi et al., 2012); grayscale image datasets USPS (Sankaranarayanan et al., 2018), and Fashion-MNIST (Xiao et al., 2017); the texture dataset DTD (Cimpoi et al., 2014); the facial emotion recognition dataset FER2013 (Goodfellow et al., 2013); the satellite image classification datasets EuroSAT (Helber et al., 2019) and RESISC45 (Cheng et al., 2017); the German Traffic Sign Recognition Benchmark (GTSRB) (Stallkamp et al., 2012); the KITTI Distance dataset (Geiger et al., 2012); the metastatic tissue classification dataset PatchCamelyon (PCam) (Veeling et al., 2018); the CLEVR counting dataset (Johnson et al., 2017); the video dataset UCF101 (Soomro et al., 2012); the multimodal HatefulMememes dataset (Kiela et al., 2020); the country classification dataset Country211 (Radford et al., 2021); and the Rendered SST2 dataset (Radford et al., 2021).

For CLEVR, we randomly sample 2,000 train and 500 test images. For UCF101, we take the middle frame of each video clip as the input. Details of each dataset are summarized in Table 2. Finally, it is worth noting that SEAL can also be applied to tasks in various modalities beyond vision, and even to cross-modal scenarios, provided that pre-trained representations are available.

Table 2: Summary statistics of the 26 datasets, including number of classes, train size, and test size.

Dataset	Num Classes	Train Size	Test Size
CLEVR (Johnson et al., 2017)	8	2,000	500
Flowers (Nilsback & Zisserman, 2008)	102	2,040	6,149
Flowers(Test) (Nilsback & Zisserman, 2008)	102	6,149	2,040
Caltech101 (Fei-Fei et al., 2004)	102	3,060	6,084
OxfordPets (Parkhi et al., 2012)	37	3,680	3,669
DTD (Cimpoi et al., 2014)	47	3,760	1,880
STL-10 (Coates et al., 2011)	10	5,000	8,000
KITTI Distance (Geiger et al., 2012)	4	5,985	1,496
USPS (Sankaranarayanan et al., 2018)	10	7,291	2,007
ImageNet-10 (Chang et al., 2017)	10	10,500	2,630
ImageNet-Dogs (Chang et al., 2017)	15	15,600	3,900
EuroSAT (Helber et al., 2019)	10	10,000	5,000
GTSRB (Stallkamp et al., 2012)	43	26,640	12,630
FER2013 (Goodfellow et al., 2013)	7	28,709	7,178
CIFAR-10 (DeVries & Taylor, 2017)	10	50,000	10,000
CIFAR-100 (DeVries & Taylor, 2017)	100	50,000	10,000
TinyImageNet (Le & Yang, 2015)	200	100,000	10,000
Fashion-MNIST (Xiao et al., 2017)	10	60,000	10,000
Food101 (Bossard et al., 2014)	101	75,750	25,250
FGVC Aircraft (Maji et al., 2013)	100	6,667	3,333
PatchCamelyon (Veeling et al., 2018)	2	294,912	32,768
UCF101 (Soomro et al., 2012)	101	9,537	3,783
Country211 (Radford et al., 2021)	211	42,200	21,100
HatefulMemes (Kiela et al., 2020)	2	8,500	500
The Rendered SST2 (Radford et al., 2021)	2	7,792	1,821
Resisc45 (Cheng et al., 2017)	45	25,200	6,300

A.2 FOUNDATION REPRESENTATIONS

SEAL is compatible with many pre-trained representations. This paper presents a comprehensive evaluation of SEAL on a wide range of representation spaces that vary in pre-training datasets, model architectures, and training objectives. Specifically, for 1-space, we consider CLIP ResNets (ResNet50, ResNet101, ResNet50x4), CLIP Vision Transformers (ViT-B/32, ViT-B/16, ViT-L/14) pre-trained on WebImageText-400M (Radford et al., 2021), as well as DINOv2 ViT-G/14 pre-trained on LVD-142M (Oquab et al., 2023). SEAL combines two representation spaces to form a 2-space. The specific backbones used in each experiment are indicated in the corresponding sections. For all models, representations are precomputed using standard image preprocessing pipelines. Details of the pre-trained representations are provided in Table 3.

Table 3: Foundation representations, including architecture, number of parameters, and textual supervision status.

Model	Architecture	Parameters	Trained on	Textual
CLIP(Radford et al., 2021)	RN50	100M	WebImageText-400M	✓
	RN101	120M		
	RN50x4	180M		
	ViT-B/32	150M		
	ViT-B/16	150M		
	ViT-L/14	430M		
DINOv2 (Oquab et al., 2023)	ViT-G/14	1.1B	LVD-142M	×

B EFFECT OF MUTUAL DISTILLATION

B.1 RESULTS OF 2-SPACE BACKBONES ON THE 26 DATASETS

To explore the benefits of mutual distillation in a 2-space setting, we evaluate SEAL on two representative 2-space backbone combinations: CLIP ViT-B/32 + CLIP ViT-B/16 and CLIP ViT-L/14 + DINOv2. Figure 8 presents the performance of SEAL (with Mutual Distillation) compared to the baseline TURTLE (without Mutual Distillation) in 26 datasets using radar charts.

For CLIP ViT-B/32 + CLIP ViT-B/16, SEAL consistently outperforms the baseline TURTLE model on most datasets. Notable improvements include datasets such as OxfordPets (+43.4%), ImageNet-Dogs (+41.41%), and EuroSAT (+10.84%), indicating that Mutual Distillation effectively improves the structural fidelity of embeddings. The radar chart illustrates that the combined representation provides more robust performance across datasets.

Similarly, for CLIP ViT-L/14 + DINOv2, SEAL demonstrates clear performance gains over TURTLE, with pronounced improvements on datasets such as GTSRB (+7.9%) and USPS(+14.3%).

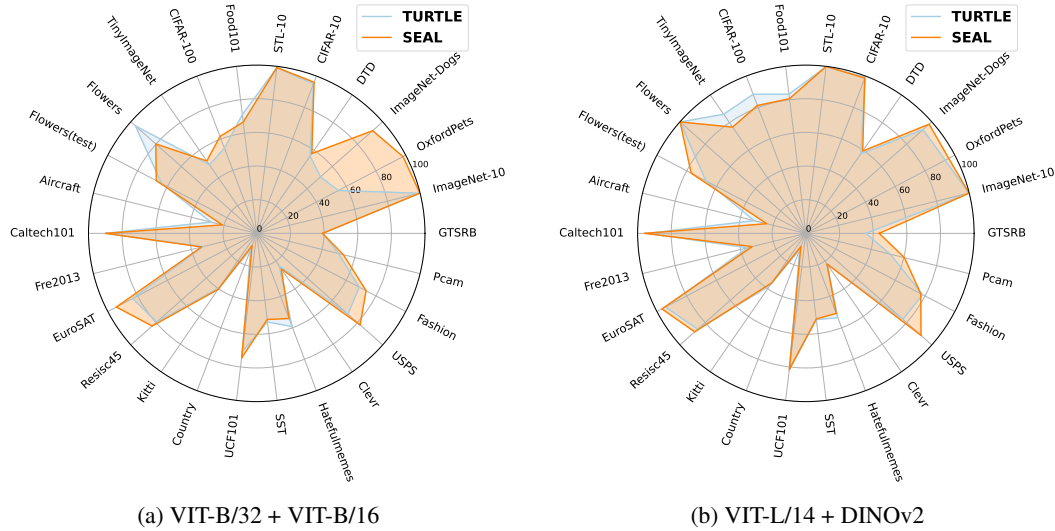


Figure 8: Effect of Mutual Distillation on 2-space backbones. The radar charts show the accuracy (%) of the TURTLE baseline (represents without Mutual Distillation) and SEAL (represents with Mutual Distillation) across 26 datasets.

B.2 RESULTS OF 1-SPACE BACKBONE ON 26 DATASETS

To evaluate the effect of SEAL on 1-space backbone, we conducted experiments on the 26 datasets. For each dataset, We measured the clustering accuracy (ACC) of SEAL (with Mutual Distillation) and the TURTLE (without Mutual Distillation) baseline across seven representative backbones, which are denoted as 1–7 in the Figure 9: (1) CLIP ResNet50, (2) CLIP ResNet101, (3) CLIP ResNet50×4, (4) CLIP ViT-B/32, (5) CLIP ViT-B/16, (6) DINOv2, and (7) CLIP ViT-L/14.

Figure 9 shows a two-panel visualization for each dataset. The upper panel displays the ACC gain ($SEAL \text{ minus } TURTLE$) for each backbone, highlighting the improvement introduced by SEAL. Points above the zero line correspond to improvements by SEAL, whereas points below indicate a decrease in performance. The lower panel directly compares the ACC of TURTLE and SEAL on per backbone. Across most of the datasets, SEAL outperforms TURTLE, achieving the largest gains on ResNet-based backbones and moderate but noticeable improvements on Transformer backbones. These results demonstrate that SEAL effectively enhances 1-space visual representations, yielding consistent and interpretable gains in clustering performance.

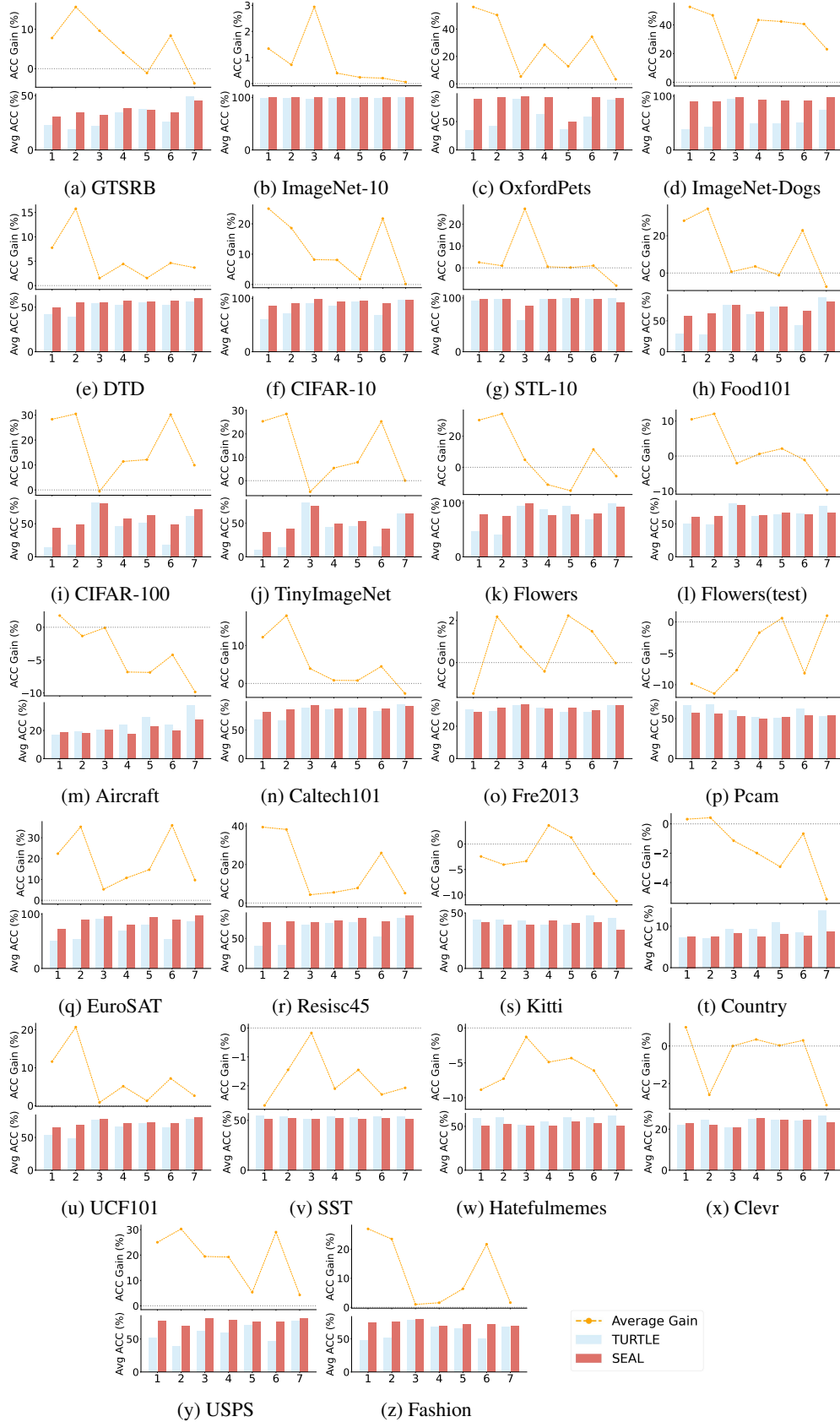


Figure 9: Effect of Mutual Distillation on 1-space backbone for each data dataset. SEAL leverages Mutual Distillation, while TURTLE serves as the counterpart without it.

C CLUSTERING STABILITY ACROSS 7 BACKBONES ON 26 DATASETS

Figure 10 presents a comprehensive view of the stability of clustering in the 26 datasets considered in our experiments. For each dataset, we evaluated the clustering accuracy distribution across 7 different backbones, comparing the performance of TURTLE and SEAL.

Overall, SEAL consistently demonstrates higher stability than TURTLE in most datasets. On datasets with simpler class structures, such as CIFAR-10, STL-10, and GTSRB, both methods achieve relatively high and stable clustering accuracy, but SEAL exhibits slightly narrower variance, indicating more reliable performance across backbones.

For more complex datasets, including ImageNet-10, OxfordPets, TinyImageNet, and Flowers, the advantage of SEAL becomes more pronounced. TURTLE’s clustering results show larger variability between different backbones, whereas SEAL maintains higher median accuracy with reduced spread, highlighting its robustness to the choice of representation.

Remote sensing datasets such as EuroSAT and Resisc45, as well as specialized datasets like Aircraft and Pcam, also benefit from SEAL’s Mutual Distillation, showing a more concentrated distribution of clustering accuracy. This indicates that SEAL can better capture fine-grained structural cues, leading to consistent clustering performance.

Finally, for more challenging datasets, such as UCF101, SEAL not only improves the median clustering accuracy but also significantly reduces the variance across backbones, demonstrating its ability to generalize well across diverse visual domains.

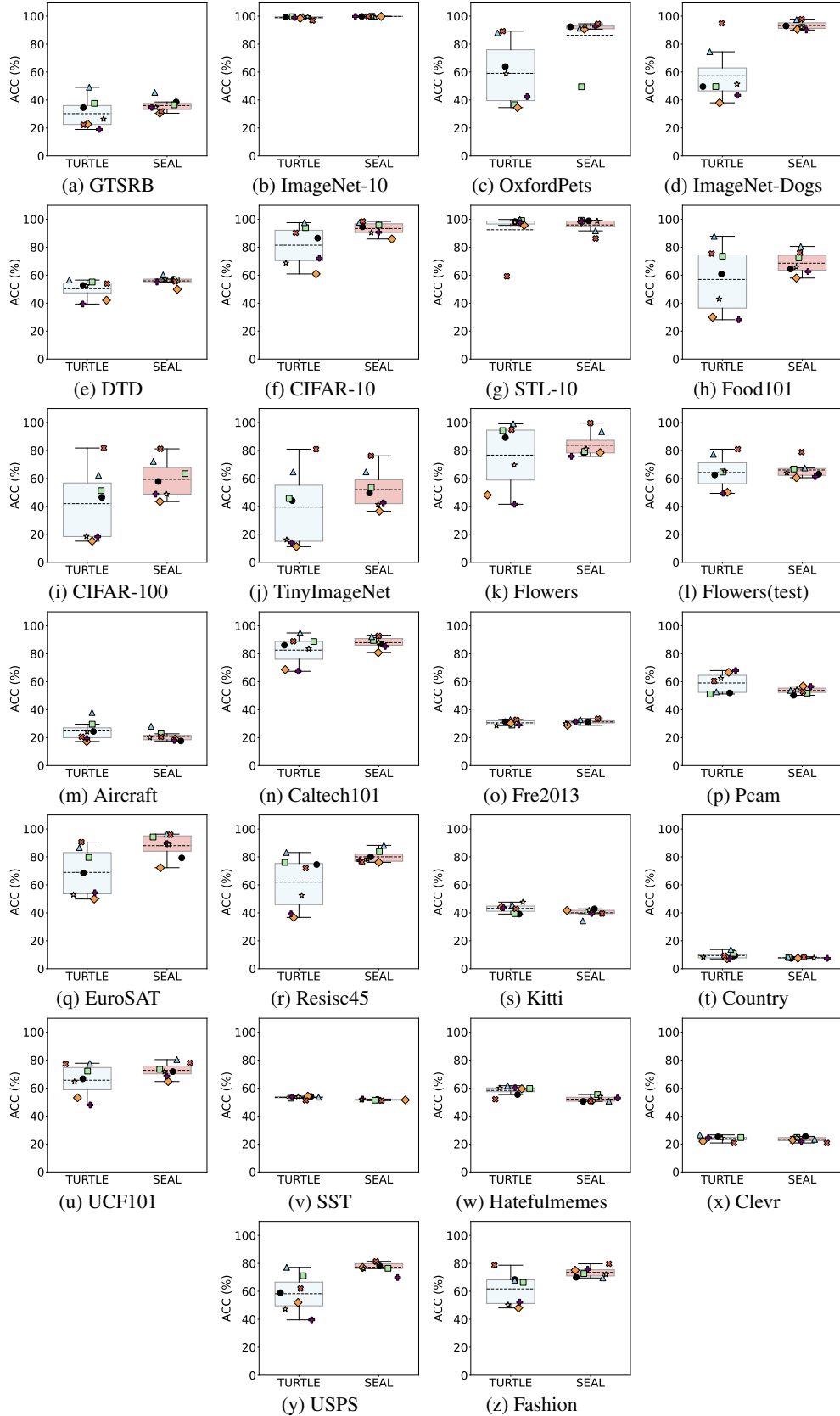


Figure 10: Detailed clustering stability plots for all 26 datasets. Each subplot shows the distribution of clustering accuracy across 7 backbones for Turtle and SEAL.

D SEPARABILITY IMPROVEMENT OF SPATIAL EMBEDDING

We evaluate the separability of representations by triplet accuracy. Triplet accuracy is computed based on the standard triplet definition: for a given anchor image, a positive image (same class) and a negative image (different class) are sampled. A triplet is considered correct if the distance between the anchor and positive embeddings is smaller than the distance between the anchor and negative embeddings. Formally, consider a triplet (x_i, x_j, x_k) , where x_i and x_j belong to the same class, and x_i and x_k belong to different classes. Let the corresponding embeddings be $f(x_i), f(x_j), f(x_k)$. The triplet is regarded as *correct* if

$$\text{dist}(f(x_i), f(x_j)) < \text{dist}(f(x_i), f(x_k)),$$

where $\text{dist}(\cdot, \cdot)$ denotes a distance function between two embeddings. In this experiment, we utilize cosine distance.

Table 4 reports the comparison of triplet verification accuracy on 10,000 randomly sampled images across 26 datasets before (*Original, the Visual Embedding*) and after (*Enhanced, the Spatially Aware Embedding*) applying mutual distillation. The results are evaluated on multiple backbone networks, including CLIP ResNets (ResNet50, ResNet101, ResNet50*4), CLIP Vision Transformers (ViT-B/32, ViT-B/16, ViT-L/14), and DINOv2.

Overall, spatially aware embedding improves triplet accuracy across almost all datasets and backbone architectures. Notable gains are observed in fine-grained datasets with high intra-class variation, such as *OxfordPets* and *ImageNet-Dogs*, demonstrating the effectiveness of incorporating spatial structure information into the foundation model embedding. On average, the triplet accuracy increases by approximately 5–6% across different backbones, with DINOv2 achieving the highest post-distillation performance.

Since higher triplet accuracy indicates that semantically similar images are closer in the embedding space while dissimilar images are farther apart, the improvement in triplet accuracy directly contributes to more separable feature clusters. Therefore, spatially aware embedding not only improves the linear separability of foundation model embeddings, making different-class data easier to distinguish, but also benefits downstream clustering performance, enabling more accurate and stable grouping of images.

E PERFORMANCE COMPARISON BETWEEN K-MEANS 1-SPACE, TURTLE 1-SPACE, AND SEAL 1-SPACE

Table 5 presents a comprehensive comparison of clustering ACC across the 26 datasets using three clustering approaches: K-Means 1-space, TURTLE 1-space, and SEAL 1-space. For each method, we evaluate multiple pre-trained backbones, including CLIP ResNets (50, 101, 50x4), CLIP ViTs (B/32, B/16, L/14) and DINOv2.

In K-Means 1-space, the image features extracted from each backbone are directly clustered using standard K-Means without additional processing. TURTLE 1-space improves on this by discovering the underlying human labeling within the feature space. SEAL 1-space further enhances performance by jointly discovering discriminative features and the underlying human labeling, effectively leveraging both spatial structure and semantic cues.

Overall, SEAL 1-space consistently achieves higher average accuracy compared to TURTLE 1-space and K-Means 1-space. The improvement is particularly notable on datasets with fine-grained categories, such as *OxfordPets*, *ImageNet-Dogs*, and *Flowers*, where SEAL 1-space maintains robust performance across different backbones.

These results indicate that SEAL captures richer information, producing more discriminative representations in a 1-space setting. This demonstrates the effectiveness of incorporating spatial and semantic cues for robust clustering across diverse datasets and backbones.

	ClipResNet50		ClipResNet101		ClipResNet50*4		ClipViT-B/32		ClipViT-B/16		DINOv2		ClipViT-L/14	
	Origin	Enhance	Origin	Enhance	Origin	Enhance	Origin	Enhance	Origin	Enhance	Origin	Enhance	Origin	Enhance
GTSRB	75.62	<u>83.99</u>	77.14	<u>85.80</u>	76.71	<u>86.43</u>	74.64	<u>82.43</u>	77.94	<u>88.38</u>	69.13	<u>82.91</u>	84.79	<u>92.90</u>
ImageNet-10	97.13	<u>99.95</u>	97.30	<u>99.93</u>	97.23	<u>99.94</u>	97.35	<u>99.92</u>	97.36	<u>99.95</u>	96.46	<u>99.92</u>	96.86	<u>99.95</u>
OxfordPets	87.93	<u>98.91</u>	91.11	<u>99.10</u>	91.20	<u>99.40</u>	88.37	<u>99.35</u>	65.14	<u>84.21</u>	83.23	<u>95.65</u>	93.29	<u>99.65</u>
ImageNet-Dogs	69.74	<u>95.94</u>	75.42	<u>97.28</u>	76.50	<u>97.62</u>	71.26	<u>97.01</u>	73.98	<u>97.30</u>	97.75	<u>99.42</u>	78.08	<u>99.05</u>
DTD	80.48	<u>89.36</u>	81.22	<u>90.11</u>	82.42	<u>90.40</u>	81.04	<u>89.65</u>	81.97	<u>90.27</u>	86.33	<u>91.06</u>	81.25	<u>91.20</u>
CIFAR-10	81.00	<u>92.87</u>	85.32	<u>95.77</u>	84.67	<u>95.66</u>	87.97	<u>97.60</u>	87.93	<u>98.45</u>	89.54	<u>99.56</u>	86.12	<u>98.93</u>
STL-10	94.26	<u>99.42</u>	94.84	<u>99.70</u>	95.06	<u>99.80</u>	94.62	<u>99.80</u>	94.28	<u>99.74</u>	83.48	<u>97.84</u>	92.14	<u>99.32</u>
Food101	89.11	<u>92.65</u>	90.65	<u>93.85</u>	91.50	<u>94.43</u>	89.70	<u>93.93</u>	92.67	<u>95.55</u>	96.95	<u>97.18</u>	94.15	<u>97.50</u>
CIFAR-100	80.07	<u>90.64</u>	84.83	<u>92.61</u>	83.67	<u>92.47</u>	86.82	<u>95.43</u>	86.44	<u>96.03</u>	96.01	<u>98.66</u>	84.74	<u>97.31</u>
TinyImageNet	84.00	<u>88.25</u>	86.94	<u>90.15</u>	86.27	<u>89.75</u>	90.68	<u>93.15</u>	90.47	<u>93.59</u>	<u>97.20</u>	95.81	89.77	<u>94.55</u>
Flowers	94.66	<u>97.21</u>	95.59	<u>96.52</u>	<u>98.04</u>	97.35	96.81	<u>96.91</u>	97.40	<u>97.75</u>	99.85	<u>99.95</u>	98.19	<u>99.46</u>
Flowers(test)	94.67	<u>97.66</u>	96.05	<u>98.10</u>	96.94	<u>98.46</u>	96.28	<u>98.32</u>	97.32	<u>98.81</u>	<u>99.89</u>	99.85	98.55	<u>99.50</u>
Aircraft	<u>79.78</u>	77.67	<u>82.30</u>	76.45	<u>84.58</u>	78.27	<u>83.05</u>	76.59	<u>86.68</u>	79.27	<u>87.48</u>	86.40	<u>89.25</u>	83.04
Caltech101	95.49	<u>98.46</u>	96.86	<u>99.22</u>	97.22	<u>98.99</u>	97.09	<u>99.15</u>	97.25	<u>99.31</u>	98.99	<u>99.61</u>	97.48	<u>99.71</u>
Fer2013	56.27	<u>59.59</u>	57.80	<u>58.91</u>	57.93	<u>59.79</u>	57.96	<u>59.18</u>	58.05	<u>59.39</u>	61.16	<u>61.42</u>	57.56	<u>61.45</u>
Pcam	<u>57.86</u>	54.19	<u>58.29</u>	53.76	<u>55.05</u>	53.31	<u>56.03</u>	52.94	<u>54.53</u>	53.28	55.41	<u>62.16</u>	<u>56.64</u>	53.90
EuroSAT	80.51	<u>94.15</u>	81.18	<u>96.27</u>	82.89	<u>96.03</u>	81.64	<u>94.56</u>	84.26	<u>97.79</u>	84.09	<u>98.01</u>	87.86	<u>98.77</u>
Resisc45	91.18	<u>96.63</u>	93.02	<u>97.49</u>	92.89	<u>97.24</u>	93.55	<u>98.01</u>	95.20	<u>98.30</u>	92.38	<u>97.35</u>	95.45	<u>98.62</u>
Kitti	59.60	<u>60.12</u>	<u>58.10</u>	55.81	59.13	<u>60.05</u>	58.48	<u>60.13</u>	58.75	<u>58.96</u>	<u>59.85</u>	59.20	<u>58.96</u>	56.91
Country	<u>64.23</u>	61.03	<u>64.73</u>	60.17	<u>66.39</u>	60.61	<u>63.64</u>	59.84	<u>65.58</u>	61.52	<u>62.55</u>	59.53	<u>67.83</u>	62.40
UCF101	92.70	<u>96.52</u>	94.63	<u>96.69</u>	94.24	<u>97.51</u>	94.57	<u>97.33</u>	95.88	<u>97.49</u>	98.22	<u>98.40</u>	97.55	<u>99.02</u>
SST	<u>50.87</u>	49.45	<u>51.63</u>	51.49	<u>52.09</u>	51.22	50.73	<u>51.57</u>	51.04	<u>51.37</u>	<u>50.73</u>	50.54	<u>52.35</u>	51.59
Hatefulmemes	<u>51.89</u>	49.81	<u>51.07</u>	50.88	50.80	<u>51.47</u>	49.61	<u>49.75</u>	50.54	<u>51.66</u>	50.14	<u>50.59</u>	50.95	<u>50.99</u>
Clevr	60.10	<u>60.10</u>	<u>62.00</u>	58.90	58.75	<u>60.75</u>	59.90	<u>61.35</u>	58.70	<u>60.55</u>	<u>55.45</u>	54.80	<u>59.20</u>	58.95
USPS	85.01	<u>96.23</u>	81.17	<u>96.80</u>	82.49	<u>96.54</u>	87.30	<u>97.83</u>	87.97	<u>97.52</u>	83.23	<u>95.65</u>	88.70	<u>97.81</u>
Fashion	81.12	<u>90.72</u>	84.10	<u>91.80</u>	82.20	<u>91.88</u>	84.82	<u>92.48</u>	82.64	<u>91.58</u>	87.48	<u>93.94</u>	81.66	<u>91.64</u>
AVG.	78.28	<u>83.52</u>	79.74	<u>83.98</u>	79.88	<u>84.77</u>	79.77	<u>84.39</u>	79.61	<u>84.54</u>	81.65	<u>85.59</u>	81.51	<u>85.39</u>

Table 4: Triplet accuracy (%) on 10,000 randomly-sampled images before (Origin) and after (Enhance) mutual distillation. Underlined numbers indicate the higher value within each Origin/Enhance pair for the same architecture.

	ClipViT-B/32			ClipViT-B/16			ClipViT-L/14			ClipResNet50			ClipResNet101			ClipResNet50*4			DINOv2		
	K-Means	TURTLE	SEAL	K-Means	TURTLE	SEAL	K-Means	TURTLE	SEAL	K-Means	TURTLE	SEAL	K-Means	TURTLE	SEAL	K-Means	TURTLE	SEAL	K-Means	TURTLE	SEAL
GTSRB	33.64	34.50	<u>38.56</u>	35.54	<u>37.59</u>	36.49	49.06	<u>49.07</u>	45.33	26.66	22.73	<u>30.52</u>	28.69	18.95	<u>34.62</u>	26.10	26.37	<u>34.77</u>	20.71	22.22	<u>31.85</u>
ImageNet-10	98.69	99.36	<u>99.76</u>	99.22	99.58	<u>99.82</u>	99.63	99.80	<u>99.86</u>	96.97	98.43	<u>99.77</u>	98.46	99.02	<u>99.74</u>	98.93	99.53	<u>99.74</u>	91.84	96.86	<u>99.80</u>
OxfordPets	52.61	63.83	<u>92.36</u>	30.38	36.63	<u>49.46</u>	69.97	88.02	<u>91.30</u>	47.77	34.54	<u>90.65</u>	57.77	42.42	<u>92.64</u>	60.57	58.75	<u>93.10</u>	82.23	89.21	<u>94.48</u>
ImageNet-Dogs	39.80	49.58	<u>92.97</u>	42.57	49.58	<u>91.94</u>	62.99	74.43	<u>97.44</u>	33.25	37.97	<u>90.51</u>	46.47	43.39	<u>90.08</u>	47.12	51.38	<u>91.96</u>	87.52	94.86	<u>97.81</u>
DTD	46.38	52.70	<u>57.13</u>	47.77	55.13	<u>56.68</u>	50.88	56.52	<u>60.21</u>	39.65	42.10	<u>49.87</u>	41.86	39.36	<u>55.13</u>	45.88	52.58	<u>57.21</u>	47.55	53.99	<u>55.51</u>
CIFAR-10	76.11	86.55	<u>94.63</u>	78.31	94.04	<u>95.81</u>	83.54	97.58	<u>97.76</u>	55.27	60.90	<u>85.90</u>	69.52	72.12	<u>90.71</u>	65.36	68.70	<u>90.42</u>	74.92	90.32	<u>98.51</u>
STL-10	97.62	98.36	<u>98.88</u>	94.84	99.12	<u>99.26</u>	95.74	<u>99.86</u>	91.72	88.94	95.66	<u>98.16</u>	96.70	97.58	<u>98.56</u>	97.02	97.72	<u>98.72</u>	53.12	59.22	<u>86.32</u>
Food101	56.13	60.86	<u>64.41</u>	70.64	<u>73.59</u>	72.42	79.80	<u>87.82</u>	80.52	<u>66.91</u>	29.93	57.97	<u>70.45</u>	28.16	62.62	<u>70.97</u>	42.98	65.88	70.65	75.43	<u>76.12</u>
CIFAR-100	42.39	46.44	<u>57.85</u>	47.82	51.31	<u>63.46</u>	53.64	62.27	<u>72.18</u>	25.04	15.13	<u>43.46</u>	34.99	18.25	<u>48.74</u>	30.36	18.56	<u>48.74</u>	68.47	80.90	<u>81.16</u>
TinyImageNet	40.85	44.15	<u>49.56</u>	43.37	45.65	<u>53.52</u>	57.49	64.66	<u>67.47</u>	25.22	11.17	<u>36.54</u>	34.40	13.97	<u>42.51</u>	34.72	16.11	<u>41.42</u>	72.19	<u>80.91</u>	76.17
Flowers	70.64	<u>89.22</u>	78.09	81.27	<u>94.26</u>	79.26	88.63	<u>99.12</u>	93.43	67.50	48.14	<u>78.43</u>	69.31	41.52	<u>75.83</u>	80.00	69.66	<u>81.13</u>	96.18	94.85	<u>99.66</u>
Flowers(test)	<u>70.64</u>	62.53	63.13	<u>81.27</u>	64.60	66.73	<u>88.63</u>	77.25	67.47	<u>67.50</u>	50.07	60.56	<u>69.31</u>	49.33	61.38	<u>80.00</u>	65.21	64.09	<u>96.18</u>	71.98	78.86
Aircraft	22.80	24.40	17.59	27.22	<u>29.58</u>	22.71	33.57	<u>37.93</u>	28.08	<u>20.25</u>	17.22	19.00	<u>23.02</u>	19.30	17.97	<u>24.31</u>	24.22	20.02	18.87	<u>20.58</u>	20.50
Caltech101	72.29	86.11	<u>86.93</u>	76.99	88.73	<u>89.51</u>	84.51	<u>94.87</u>	92.22	62.06	68.63	<u>80.82</u>	69.22	67.42	<u>85.29</u>	73.24	83.56	<u>88.04</u>	85.82	88.89	<u>92.81</u>
Fer2013	26.95	<u>31.38</u>	30.96	27.82	28.99	<u>31.22</u>	31.48	<u>32.93</u>	32.91	26.42	<u>30.28</u>	28.81	26.79	29.12	<u>31.30</u>	27.65	28.58	<u>30.07</u>	32.98	32.81	<u>33.56</u>
Pcam	<u>62.05</u>	51.97	50.27	<u>62.24</u>	51.22	51.83	<u>61.32</u>	52.73	53.73	64.30	<u>66.68</u>	56.84	64.28	<u>67.93</u>	56.53	62.01	<u>62.36</u>	54.19	59.43	<u>60.50</u>	52.83
EuroSAT	64.14	68.55	<u>79.28</u>	73.98	79.59	<u>94.27</u>	74.99	86.60	<u>96.29</u>	56.83	49.89	<u>72.27</u>	52.31	54.31	<u>89.54</u>	53.39	52.93	<u>88.89</u>	62.36	60.61	<u>95.89</u>
Resisc45	68.19	74.56	<u>80.07</u>	73.36	76.05	<u>83.83</u>	73.87	83.14	<u>88.27</u>	54.29	36.72	<u>76.12</u>	62.56	39.27	<u>77.45</u>	66.54	52.44	<u>78.43</u>	62.65	71.98	<u>76.30</u>
Kitti	<u>48.09</u>	39.13	42.79	<u>48.89</u>	39.43	40.70	<u>48.35</u>	45.45	34.25	<u>47.25</u>	44.11	41.67	<u>48.32</u>	39.55	39.55	<u>47.95</u>	41.82	41.82	<u>49.16</u>	42.79	39.43
Country	<u>9.51</u>	9.38	7.39	10.57	<u>10.98</u>	8.06	12.99	<u>13.74</u>	8.60	<u>8.64</u>	7.22	7.54	<u>9.37</u>	7.03	7.45	<u>9.87</u>	8.39	7.72	9.23	<u>9.34</u>	8.19
UCF101	61.34	66.66	<u>71.29</u>	63.53	72.15	<u>73.44</u>	70.94	77.75	<u>80.38</u>	51.87	53.09	<u>64.71</u>	59.64	47.89	<u>68.59</u>	61.85	64.54	<u>71.72</u>	71.33	77.26	<u>78.10</u>
SST	52.21	<u>54.03</u>	51.93	<u>54.52</u>	52.77	51.32	<u>51.73</u>	53.49	51.42	51.75	<u>54.18</u>	51.50	<u>53.80</u>	53.62	52.17	<u>55.25</u>	53.93	53.95	51.16	<u>51.21</u>	51.04
Hatefulmemes	<u>60.15</u>	55.38	50.49	<u>60.91</u>	59.74	55.42	<u>62.72</u>	61.71	50.60	<u>60.85</u>	59.48	50.62	<u>60.59</u>	60.20	52.93	<u>60.86</u>	60.05	54.19	<u>55.60</u>	52.00	50.72
Clevr	24.95	25.10	<u>25.45</u>	24.30	24.60	<u>24.63</u>	23.50	<u>26.50</u>	23.35	24.05	21.90	<u>22.90</u>	<u>26.05</u>	24.50	21.90	<u>25.30</u>	24.25	24.55	18.70	20.80	<u>20.80</u>
USPS	60.76	59.00	<u>78.23</u>	78.47	71.13	<u>76.49</u>	79.66	77.16	<u>81.44</u>	61.42	52.00	<u>77.04</u>	51.28	39.57	<u>69.87</u>	51.83	47.25	<u>76.30</u>	56.38	61.97	<u>81.42</u>
Fashion	64.16	68.43	<u>70.01</u>	58.96	66.31	<u>72.69</u>	53.19	67.93	<u>69.56</u>	58.78	48.10	<u>75.09</u>	63.83	52.27	<u>75.77</u>	56.05	50.29	<u>72.00</u>	65.37	78.71	<u>79.72</u>
AVG.	54.75	57.78	<u>62.01</u>	57.23	59.71	<u>63.11</u>	62.79	<u>68.01</u>	67.42	48.64	44.47	<u>59.59</u>	52.71	45.00	<u>61.50</u>	53.43	51.08	<u>62.56</u>	59.93	64.61	<u>67.60</u>

F PERFORMANCE COMPARISON BETWEEN K-MEANS 2-SPACE, TURTLE 2-SPACE, AND SEAL 2- SPACE

Table 6 presents a detailed comparison of clustering accuracy across 26 diverse vision datasets using three 2-space clustering approaches: K-Means 2-space, TURTLE 2-space, and SEAL 2-space. Each method is evaluated with different backbone combinations, including CLIP VIT-L/14 + DINOv2 and CLIP VIT-B/32 + CLIP VIT-B/16.

Overall, SEAL 2-space achieves the highest average accuracy compared to TURTLE 2-space and K-Means 2-space. The performance gains are particularly notable on datasets with complex structures categories, such as EuroSAT, CIFAR-10, OxfordPets, and ImageNet-Dogs, where SEAL 2-space maintains robust performance across different combinations.

In addition, we report the clustering performance of K-means on the 26 datasets in two settings, which are (1) directly concatenating the CLIP VIT-L/14 and DINOv2 embeddings and (2) applying the mutual distillation to both embeddings before concatenation. The results are listed in Table 7. From Table 7, it can be seen that spatially aware embeddings significantly improve performance on most of the datasets, which illustrates the effectiveness of mutual distillation.

These results demonstrate that jointly leveraging two embedding spaces and spatial structure cues enables SEAL to produce more discriminative and semantically meaningful representations, leading to superior clustering performance compared to standard K-Means and TURTLE in a 2-space setting.

	VIT-B/32 / VIT-B/16			VIT-L/14 / DINOv2		
	K-Means	TURTLE	SEAL	K-Means	TURTLE	SEAL
GTSRB	36.25	<u>39.33</u>	39.09	22.81	35.73	<u>43.61</u>
ImageNet-10	99.18	99.62	<u>99.82</u>	92.67	99.79	<u>99.87</u>
OxfordPets	52.69	54.81	<u>98.18</u>	83.10	93.54	<u>95.87</u>
ImageNet-Dogs	42.30	50.62	<u>92.03</u>	86.37	93.03	<u>97.87</u>
DTD	49.04	55.85	<u>57.61</u>	48.88	57.45	<u>59.44</u>
CIFAR-10	79.14	95.01	<u>96.07</u>	85.00	<u>99.47</u>	98.69
STL-10	98.34	99.22	<u>99.26</u>	51.54	99.74	<u>99.94</u>
Food101	66.94	<u>71.08</u>	66.63	72.08	<u>83.31</u>	80.63
CIFAR-100	48.60	54.08	<u>61.59</u>	69.72	<u>88.38</u>	81.28
TinyImageNet	44.65	49.55	<u>52.27</u>	74.70	<u>86.02</u>	76.70
Flowers	82.35	<u>96.27</u>	80.15	98.68	99.80	<u>99.85</u>
Flowers(test)	<u>74.58</u>	63.95	67.43	<u>93.71</u>	67.12	77.00
Aircraft	26.73	<u>28.09</u>	21.06	18.75	<u>31.66</u>	23.98
Caltech101	75.33	89.77	<u>89.87</u>	85.46	93.86	<u>95.88</u>
Fre2013	27.74	32.87	<u>33.76</u>	33.10	<u>35.76</u>	32.89
Pcam	<u>67.20</u>	50.03	53.00	58.76	51.67	<u>60.33</u>
EuroSAT	71.85	83.37	<u>94.21</u>	62.72	94.20	<u>96.60</u>
Resisc45	49.06	79.38	<u>83.04</u>	67.38	85.90	<u>88.27</u>
Kitti	9.95	39.31	<u>40.47</u>	<u>49.16</u>	37.94	35.69
Country	<u>65.32</u>	10.21	7.84	9.28	<u>10.25</u>	8.36
UCF101	54.94	70.67	<u>74.57</u>	72.47	<u>81.81</u>	81.38
SST	<u>60.62</u>	53.04	51.55	51.17	<u>51.66</u>	51.23
Hatefulmemes	24.95	<u>59.25</u>	54.13	<u>55.67</u>	53.59	50.89
Clevr	<u>82.47</u>	23.45	25.65	20.05	22.10	22.10
USPS	65.67	72.42	<u>82.03</u>	57.66	77.11	<u>91.43</u>
Fashion	62.71	68.81	<u>73.40</u>	65.38	<u>78.33</u>	77.09
AVG.	58.41	61.16	<u>65.18</u>	61.01	69.59	<u>70.26</u>

Table 6: Clustering accuracy comparison across different backbone combination settings and methods. The Underlined value indicates the best-performing method within that combination setting.

Dataset	K-Means Clustering on	
	Directly Concatenating ViT-L/14 and DINOv2	Spatially Aware Embedding with ViT-L/14 and DINOv2
GTSRB	22.81	<u>41.14</u>
ImageNet-10	92.67	<u>99.79</u>
OxfordPets	83.10	<u>90.41</u>
ImageNet-Dogs	86.37	<u>97.66</u>
DTD	48.88	<u>60.00</u>
CIFAR-10	85.00	<u>98.66</u>
STL-10	51.54	<u>99.56</u>
Food101	72.08	<u>75.51</u>
CIFAR-100	69.72	<u>72.57</u>
TinyImageNet	<u>74.70</u>	70.54
Flowers	<u>98.68</u>	94.90
Flowers(test)	<u>93.71</u>	91.38
Aircraft	18.75	<u>21.84</u>
Caltech101	85.46	<u>88.27</u>
Fre2013	33.10	<u>33.85</u>
Pcam	58.76	<u>60.21</u>
EuroSAT	62.72	<u>96.26</u>
Resisc45	67.38	<u>79.03</u>
Kitti	<u>49.16</u>	46.68
Country	<u>9.28</u>	8.18
UCF101	72.47	<u>78.83</u>
SST	51.17	<u>51.18</u>
Hatefulmemes	<u>55.67</u>	50.73
Clevr	20.05	<u>20.35</u>
USPS	57.66	<u>82.79</u>
Fashion	65.38	<u>71.48</u>
AVG.	61.01	<u>68.53</u>

Table 7: K-Means accuracy from concatenating embeddings of ViT-L/14 and DINOv2 after applying a spatially-aware transformation to each, compared to directly concatenating original embeddings of ViT-L/14 and DINOv2, on 26 benchmark datasets.