EVR: Eyesight-Only Visual Representation for Recognition without Language

Anonymous CVPR submission

Paper ID *****

Abstract

We present EVR, a Visual Self-Consistent Representation learning framework that builds a semantic space solely from visual signals. EVR constructs a structured set of visual prototypes by clustering rich visual features and models the relations among prototypes with a graph neural architecture. A self-supervised semantic discovery module refines prototypes and image encoders jointly through prototype-aware contrastive clustering objectives, enabling the system to discover intermediate-granularity semantics without any textual supervision. Such semantics are more abstract than individual instances yet remain subordinate to manually designated categories. When human-readable labels are required for downstream interpretation, an optional decoding step uses a language model to map learned prototypes to pseudo-labels, but the training pipeline remains entirely text-free. EVR reduces language-induced biases, improves robustness for dense and open-vocabulary tasks, and offers a practical alternative to approaches that rely on large-scale image-text pairs.

Keywords: Representation Learning; Visual prototypes; Prototype graph; Self-supervision; Open-vocab segmentation

1. Introduction

Large-scale vision-language systems that align images with text have demonstrated impressive transfer capabilities across a wide range of visual tasks, but their reliance on textual supervision can introduce language-driven biases, translation ambiguities, and fragility under domain shift [1, 9, 36, 51, 61, 71]. Techniques that attach prompts or lightweight adapters to frozen vision-language backbones improve adaptability, yet these approaches remain text-centric and typically presuppose reliable textual anchors. In parallel, an increasing body of work shows that strong visual pretraining combined with prototype-oriented clustering is capable of recovering meaningful visual semantics without text [50, 56, 70–72]. Such purely visual strategies reveal category-like clusters and meso-level structure, which suggests that a visual-only semantic space can be both expressive and broadly transferable.

Despite these promising developments, two interconnected limitations persist. First, many vision-language pipelines fuse text and image signals early or treat language as the primary semantic scaffold, which risks overwriting general visual features with dataset-specific textual idiosyncrasies and thereby degrading zero-shot and open-vocabulary generalization [10, 61]. Second, a large fraction of prototype-based visual methods either assign a single prototype to each concept or neglect explicit modeling of prototype-to-prototype relations. This restricts the capacity to represent intra-class diversity and to capture contextual interactions among prototypes that are important for dense prediction and compositional recognition [9, 51, 78]. Consequently, current visual-only and hybrid solutions often trade interpretability, robustness, and open-vocabulary transferability against one another. To address these gaps, we introduce EVR, a visual-first framework that constructs a self-consistent semantic space from images alone. It consists of three components: the Visual Prototype Graph partitions the feature manifold into semantic clusters and models prototype relations via learned message passing; the Self-Supervised Semantic Discovery module jointly updates prototypes and encoder parameters using contrastive clustering with consistency regularizers [50, 56, 70]; and a lightweight textual decoder provides post-hoc interpretability without influencing the learned visual space.

EVR differs from prior prototype and self-supervised pipelines by treating visual prototypes as primary semantic units and explicitly modeling their relations via learned graph propagation. Unlike methods such as SwAV[3], DINO[4], and ProtoNCE[26] that rely on single-centroid or instance-level grouping without relational structure, EVR connects multiple prototypes per concept to capture intra-class diversity, contextual co-occurrence, and compositional interactions. By decoupling semantic learning from language, EVR avoids early textual bias and improves robustness in open-vocabulary and dense prediction tasks [1, 36, 53]. It builds a purely visual semantic manifold with multi-prototype modeling, adaptive loss normalization, and

stability diagnostics to balance smoothing and discrimination [34, 78].

This paper makes the following contributions. We present EVR, a framework that constructs a structured, purely visual semantic space by combining prototype clustering, graph-based prototype modeling, and self-supervised semantic discovery, thereby removing the need for text during representation learning. We introduce the Visual Prototype Graph architecture and show how graph-based propagation captures inter-prototype dependencies and supports multi-prototype modeling to represent intra-class diversity. We propose a Self-Supervised Semantic Discovery objective that couples prototype refinement and encoder updates through prototype-aware contrastive clustering and consistency regularization, enabling robust prototype emergence without language supervision. Finally, we empirically demonstrate that EVR improves robustness and transfer in dense prediction and open-vocabulary scenarios, and we analyze how prototype structure, propagation, and the proposed stability mechanisms reduce language-induced artifacts relative to vision-language baselines and prior prototype approaches. The experiments and ablations reported below validate EVR's effectiveness and highlight practical engineering defaults for reproducible training at scale.

2. Related Work

This section situates our work within three tightly related areas: adaptation and tuning of vision-language models, prototype- and cluster-based visual representation learning, and graph-based semantic propagation and stability techniques. We emphasize representative developments and practical engineering practices that have guided the design choices in EVR.

2.1. Adaptation and Tuning of Vision-Language Models

Recent work explores adapting large vision-language models (VLMs) to downstream tasks with limited supervision. Strategies include prompt tuning, parameter-efficient adapters, low-rank updates, and semantically guided visual adaptation [43, 48, 69]. Studies highlight sensitivity to hyperparameters and validation protocols, prompting robust tuning methods and two-stage pipelines [12, 38, 48]. Domain-specific methods, such as in medical imaging, leverage priors or lightweight tuning for efficiency [42, 47]. Other approaches use online or meta-learned adapters for continual or few-shot updates [49]. In contrast, our method constructs a purely visual semantic space using prototype-based adaptation without textual supervision [11].

2.2. Prototype-based self-supervised representation learning

Discovering and leveraging prototype representations via self-supervised clustering has emerged as an effective strategy for unsupervised semantic discovery and downstream generalization. Prior works propose prototype discovery routines, prototype-augmented generators, and prototype-centered training objectives that mitigate assignment collapse and reveal novel categories [7, 57, 71]. Practical measures such as entropy regularization, buffer-based reinitialization, class-balanced sampling and EMA updates have been used to stabilize prototype optimization and to reduce the incidence of inactive or "ghost" prototypes in long-tail settings [32, 68, 73]. Our method builds on these ideas and integrates targeted monitoring and reinitialization mechanisms to maintain prototype utilization across training.

2.3. Graph-structured prototype propagation and relational modeling

Graph-based structures enable modeling of higher-order and contextual relations beyond pairwise similarity [27, 55, 58]. To address oversmoothing and heterophily issues [22, 33, 45], EVR applies prototype graph propagation with adaptive normalization and tension-aware control, balancing smoothing and separation.

2.4. Training stability, loss normalization and engineering practices

Training stability is critical in self-supervised and prototype-based systems. Prior work recommends loss normalization, gradient clipping, EMA updates, and principled hyperparameter tuning to prevent instability [7, 29, 74]. In VLM adaptation, reproducibility across datasets and scales is emphasized through sensitivity analyses and ablations [11, 48]. Following these practices, EVR adopts default settings for EMA decay, running-window estimation, buffer sizes, and reinitialization heuristics to ensure stable and reproducible training.

2.5. Applications, extensions and related paradigms

Prototype- and graph-centric ideas have found applications across zero-shot recognition, class-incremental learning, remote sensing, medical imaging, and tracking tasks [32, 42, 57, 67]. Other contemporaneous lines of work explore dynamic prototype learning within multimodal models, distribution-aware prompt tuning, and graph adapters for VLM fine-tuning [5, 28, 81]. The modularity of prototype graphs makes them amenable to these extensions, and our proposal is designed to be interoperable with such paradigms while preserving a language-free training protocol.

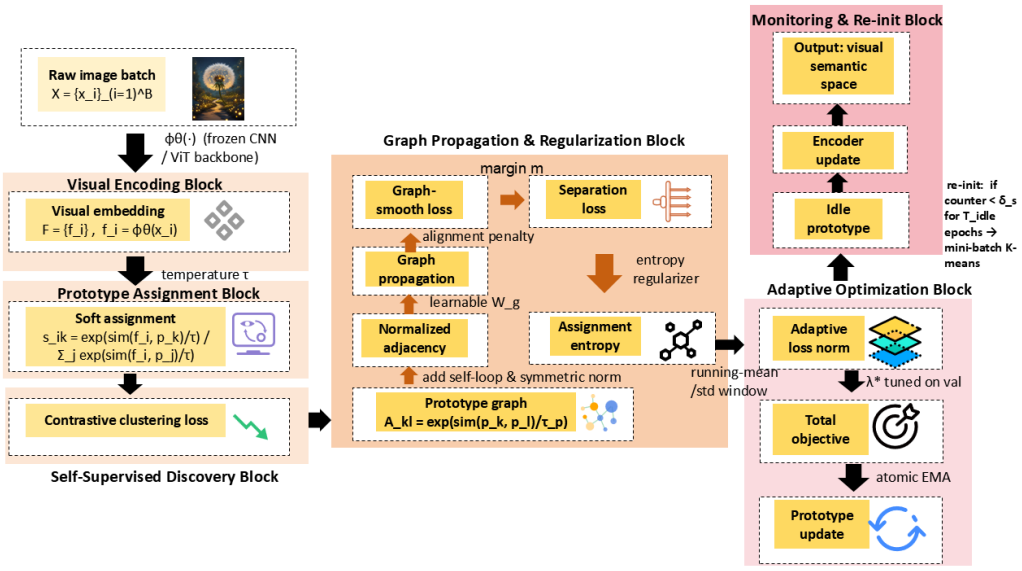

Figure 1. Overview of the EVR framework. A frozen encoder extracts image features, which are softly assigned to learnable visual prototypes. These prototypes are refined via contrastive clustering and graph-based propagation, capturing semantic relations without textual supervision. Regularization and EMA updates ensure stability, while idle prototypes are recycled to maintain diversity. An optional language decoder can be attached at inference for interpretability.

3. Methodology

This section introduces VSCR, the core of EVR, which learns semantic manifolds purely from visual signals. It discovers pseudo-semantic prototypes via self-supervised clustering, models their relations through graph structures, and jointly optimizes embeddings and prototypes under robust constraints.

3.1. Preliminaries and notation

Let $\mathcal{X} = \{x_i\}_{i=1}^N$ denote the training image set and let $\phi_\theta : \mathcal{X} \to \mathbb{R}^d$ be the visual encoder parameterized by θ. For an image x_i we define its d-dimensional embedding as

$$f_i = \phi_\theta(x_i), \tag{1}$$

where $f_i \in \mathbb{R}^d$ and θ denotes encoder parameters.

We maintain K learnable visual prototypes arranged in the prototype matrix

$$\mathcal{P} = \{p_k\}_{k=1}^K, \qquad P = [p_1^\top; \dots; p_K^\top] \in \mathbb{R}^{K \times d}, \tag{2}$$

where each prototype $p_k \in \mathbb{R}^d$ represents a pseudo-semantic center.

Similarity between vectors u and v is measured by cosine similarity:

$$\mathrm{sim}(u, v) = \frac{u^\top v}{\|u\|\|v\|}, \tag{3}$$

where $\|\cdot\|$ denotes the Euclidean norm.

3.2. Soft assignment and contrastive clustering

Instance-to-prototype soft assignments are computed with a temperature-scaled softmax over cosine similarities:

$$s_{ik} = \frac{\exp\left(\mathrm{sim}(f_i, p_k)/\tau\right)}{\sum_{j=1}^K \exp\left(\mathrm{sim}(f_i, p_j)/\tau\right)}. \tag{4}$$

where $s_{ik} \in (0, 1)$ is the soft assignment of embedding f_i to prototype p_k, K is the number of prototypes and $\tau > 0$ is the assignment temperature.

Given two stochastic augmentations of the same image that produce embeddings f_i and $\tilde{f}_i$, we enforce cross-view agreement with a symmetric contrastive-clustering loss:

$$\mathcal{L}_{\mathrm{cc}} = -\frac{1}{2N} \sum_{i=1}^N \sum_{k=1}^K \Big(s_{ik} \log \frac{\exp\left(\mathrm{sim}(f_i, p_k)/\tau\right)}{\sum_{j=1}^K \exp\left(\mathrm{sim}(f_i, p_j)/\tau\right)} + \tilde{s}_{ik} \log \frac{\exp\left(\mathrm{sim}(\tilde{f}_i, p_k)/\tau\right)}{\sum_{j=1}^K \exp\left(\mathrm{sim}(\tilde{f}_i, p_j)/\tau\right)} \Big). \tag{5}$$

where $\tilde{s}_{ik}$ denotes the assignment for the second augmented view $\tilde{f}_i$ and $\mathcal{L}_{\mathrm{cc}}$ encourages cross-view prototype consistency while mitigating collapse. To decouple prototype updates from assignments, EVR supports two options: using detached prototypes during assignment or applying EMA updates. The recommended choice depends on the training setup.

3.3. Prototype graph construction and normalization

High-order semantic structure among prototypes is captured by constructing an undirected prototype graph whose raw adjacency is given by

$$A_{kl} = \exp\big(\mathrm{sim}(p_k, p_l)/\tau_p\big), \tag{6}$$

where $\tau_p > 0$ is the graph temperature that controls adjacency sparsity.

We include self-loops and symmetrically normalize the adjacency matrix before propagation:

$$\widetilde{A} = A + I, \tag{7}$$

$$\widetilde{D} = \mathrm{diag}(\widetilde{A}\mathbf{1}), \qquad \widehat{A} = \widetilde{D}^{-\frac{1}{2}}\,\widetilde{A}\,\widetilde{D}^{-\frac{1}{2}}, \tag{8}$$

where I is the identity matrix, $\mathbf{1}$ is the all-ones vector and $\widetilde{D}$ is the degree matrix. For numerical robustness we clamp diagonal entries of $\widetilde{D}$ by a small constant $\epsilon > 0$.

3.4. Prototype propagation and graph-smoothness regularization

We propagate prototypes by a single linear transform over the normalized adjacency:

$$P' = \widehat{A}\,P\,W_g, \tag{9}$$

where $W_g \in \mathbb{R}^{d\times d}$ is a learnable linear mapping and P' denotes the propagated prototype matrix.

To encourage geometric alignment between original and propagated prototypes we introduce a graph-smoothing penalty:

$$\mathcal{L}_{\mathrm{gs}} = \frac{1}{K}\sum_{k=1}^{K}\big(1 - \mathrm{sim}(p_k, p'_k)\big), \tag{10}$$

where p'_k denotes the k-th row of P' and $\mathcal{L}_{\mathrm{gs}}$ penalizes inconsistencies created by propagation.

3.5. Prototype separation and assignment entropy

To avoid prototype collapse and encourage inter-prototype discrimination we adopt a margin-based separation loss:

$$\mathcal{L}_{\mathrm{sep}} = \frac{1}{K(K-1)}\sum_{k\neq l}\max\big(0, m - \mathrm{sim}(p_k, p_l)\big), \tag{11}$$

where $m \in (0, 1]$ is a separation margin and positive m pushes highly similar prototypes apart.

We also penalize low-entropy assignments by introducing an assignment entropy regularizer:

$$\mathcal{L}_{\mathrm{ent}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} s_{ik}\log s_{ik}, \tag{12}$$

where $\mathcal{L}_{\mathrm{ent}}$ encourages balanced prototype utilization and reduces the incidence of error prototypes.

3.6. Adaptive Loss Normalization and Total Objective

To balance loss terms with varying scales, each component is normalized using its running mean and standard deviation over a recent window of accumulated samples. Let $\bar{\mathcal{L}}_{\mathrm{cc}}, \bar{\mathcal{L}}_{\mathrm{gs}}, \bar{\mathcal{L}}_{\mathrm{sep}}, \bar{\mathcal{L}}_{\mathrm{ent}}$ denote the normalized losses. The total objective is:

$$\mathcal{L} = \lambda_{\mathrm{cc}}\bar{\mathcal{L}}_{\mathrm{cc}} + \lambda_{\mathrm{gs}}\bar{\mathcal{L}}_{\mathrm{gs}} + \lambda_{\mathrm{sep}}\bar{\mathcal{L}}_{\mathrm{sep}} + \lambda_{\mathrm{ent}}\bar{\mathcal{L}}_{\mathrm{ent}}, \tag{13}$$

where $\lambda_* \geq 0$ are scalar weights. The default window size corresponds to approximately 12.8k accumulated samples to ensure consistency under gradient accumulation.

3.7. Stability Mechanisms, Prototype Monitoring and Reinitialization

EVR ensures training stability through three mechanisms. First, prototypes are updated via an atomic exponential moving average (EMA) within the optimizer step:

$$p_k^{(t+1)} \leftarrow \alpha_{\mathrm{ema}}\,p_k^{(t)} + (1 - \alpha_{\mathrm{ema}})\,\hat{p}_k^{(t)}, \tag{14}$$

where $p_k^{(t)}$ is the current prototype, $\hat{p}_k^{(t)}$ is the gradient-updated value, and $\alpha_{\mathrm{ema}} = 0.996$ controls the update rate.

Second, a memory-efficient buffer stores recent embeddings for prototype reinitialization. For large K, clustering is performed on a reduced buffer ($K_{\mathrm{buf}} = 2000$) to select representative samples. Third, prototype usage is tracked via atomic counters. A prototype is reinitialized using mini-batch K-means if its assignment mass stays below $\delta_s = 10^{-3}$ for $T_{\mathrm{idle}} = 10$ epochs. Other defaults include $\epsilon = 10^{-6}$ and $G_{\max} = 1.0$.

3.8. Practical rules of thumb, automatic tension control and switching rule

To prevent adjacency oscillation, we bind τ_p to τ via a ratio:

$$\tau_p \leftarrow r_\tau \cdot \tau, \qquad r_\tau = 0.5 \tag{15}$$

and increase r_τ to 0.6 when utilized prototypes fall below 50% of K.

To balance opposing forces from $\mathcal{L}_{\mathrm{gs}}$ and $\mathcal{L}_{\mathrm{sep}}$, we define a tension indicator:

$$T = \frac{1}{\binom{K}{2}}\sum_{k\neq l}\mathrm{sim}(p_k, p_l) \;-\; \frac{1}{K}\sum_{k}\mathrm{sim}(p_k, p'_k). \tag{16}$$

Weights are adjusted when T exceeds bounds (e.g., $T > 0.8$ or $T < 0.2$). For training-mode switching, we use a hybrid rule: if the fraction of tail classes with fewer than $n_{\mathrm{abs}} = 5$ samples exceeds $r_{\mathrm{tail}} = 0.05$, we apply prototype-tuning; otherwise, joint fine-tuning is preferred.

3.9. Inference and optional post-hoc labeling

At inference, an input x is embedded as $f = \phi_\theta(x)$, and its soft assignment $s(f) = (s_1(f), \ldots, s_K(f))$ is computed via Eq. (4). Prediction uses either the argmax index $\hat{k} = \arg\max_k s_k(f)$ or the full distribution for calibrated outputs. Mapping prototypes to human-readable labels is optional and done post-hoc using a small labeled seed set or external language models. No text supervision is used during training.

3.10. Algorithm: EVR Training Loop

Algorithm 1 summarizes the EVR training process, which combines dual-view contrastive clustering, loss normalization, and atomic EMA updates for stability. It also includes memory-efficient buffering, prototype usage tracking, and a dual-threshold strategy for switching between pseudo-target and joint feature-target modes.

4. Experiments

We evaluate EVR using four widely adopted protocols in the vision-language adaptation literature: base-to-novel generalization, cross-dataset transfer, domain generalization, and few-shot learning. All experiments, unless otherwise stated, are conducted under a 16-shot training regime, meaning each class is provided with 16 labeled examples. This setup ensures a fair comparison with previous methods.

4.1. Visualization Results

This section presents key visualizations illustrating EVR's training dynamics, prototype behavior, and component effectiveness.

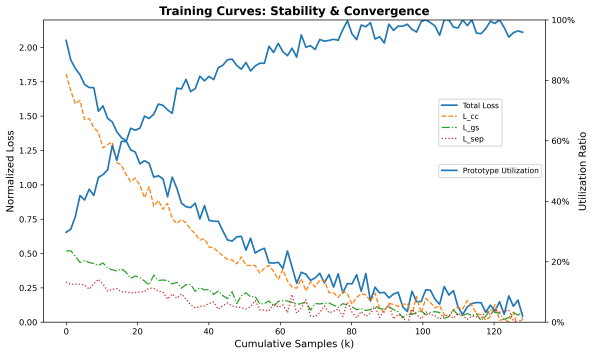

Figure 2. Training curves showing stable convergence and consistent prototype utilization enabled by loss normalization and EMA updates.

4.2. Datasets and Evaluation Protocols

We follow standard benchmarks used in prompt- and adapter-based methods. The main datasets include ImageNet [8], Caltech101 [13], OxfordPets [41], StanfordCars [25], Flowers102 [39], Food101 [2], FGVCAircraft [37], SUN397 [59], UCF101 [39], DTD [6], and EuroSAT [18]. For domain generalization, we use ImageNetV2 [46], ImageNet-Sketch [54], ImageNet-A [20], and ImageNet-R [19]. In base-to-novel evaluation, classes are split into base and novel sets. Models are trained on base labels and evaluated on both. We report top-1 accuracy and

Algorithm 1 EVR training loop

Require: dataset $\{x_i\}_{i=1}^N$, encoder ϕ_θ, prototypes P, graph transform W_g, hyperparameters $\{K, \tau, r_\tau, \lambda_*, m, \epsilon, \alpha_{\text{ema}}, G_{\max}, \delta_s,$
Require: $T_{\text{idle}}, n_{\text{abs}}, r_{\text{tail}}, E_{\text{warm}}, K_{\text{buf}}\}$

for each epoch **do**
  for each mini-batch $\mathcal{B}$ **do**
    sample two augmentations per image and compute embeddings $\{f_i, \tilde{f}_i\}_{i \in \mathcal{B}}$
    compute soft assignments s_{ik} and $\tilde{s}_{ik}$ using Eq. (4); optionally compute assignments with prototypes detached
    compute contrastive-clustering loss $\mathcal{L}_{\text{cc}}$ via Eq. (5)
    form adjacency A via Eq. (6), add self-loops $\widetilde{A} = A + I$, compute $\widetilde{D}$ and clamp $\widetilde{D}_{ii} \geq \epsilon$
    form normalized adjacency $\widehat{A}$ via Eq. (8) and compute $P' = \widehat{A} P W_g$
    compute $\mathcal{L}_{\text{gs}}, \mathcal{L}_{\text{sep}}, \mathcal{L}_{\text{ent}}$ using Eqs. (10),(11),(12)
    normalize each loss by running statistics counted on cumulative-sample windows and form total loss $\mathcal{L}$ via Eq. (13)
    compute gradients via backprop
    perform atomic EMA prototype update in-place for all k:
    p_k.data $\leftarrow \alpha_{\text{ema}} p_k$.data $+ (1 - \alpha_{\text{ema}})\, \hat{p}_k$
    (this update is performed before optimizer.step() to avoid a one-step lag)
    apply gradient clipping with max norm $G_{\max}$ and **call optimizer.step()** to update non-prototype parameters
    atomically increment per-prototype utilization counters for assignments $s_{ik} > \delta_s$
    at epoch end, detect idle prototypes (counters $<$ threshold) and reinitialize idle prototypes by mini-batch K-means on recent buffer (clustered buffer if $K > 5$k)
    monitor tension T using Eq. (16) and adjust $(\lambda_{\text{sep}}, \lambda_{\text{gs}})$ when T falls outside target band
  end for
  reset per-prototype counters for next epoch
end for

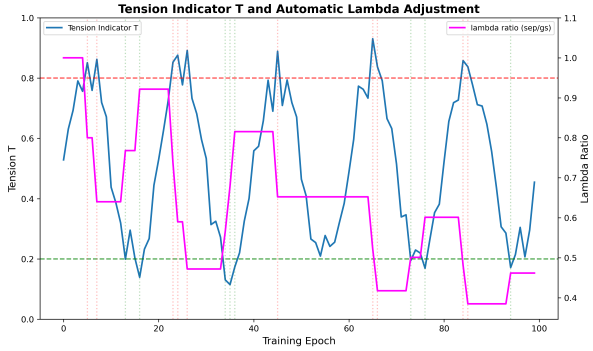

Figure 3. Tension indicator T and adaptive λ ratio dynamics. Thresholds trigger automatic weight adjustments during training.

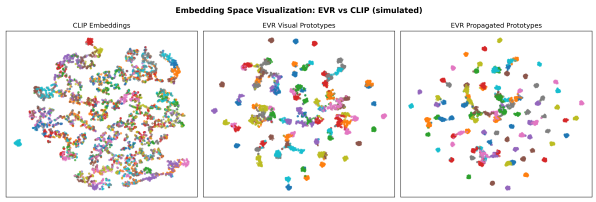

Figure 4. t-SNE visualization of EVR vs. CLIP embeddings on ImageNet. EVR shows tighter intra-class clustering and better inter-class separation.

harmonic mean (HM):

$$\mathrm{HM} = \frac{2 \cdot \mathrm{Acc}_{\mathrm{base}} \cdot \mathrm{Acc}_{\mathrm{novel}}}{\mathrm{Acc}_{\mathrm{base}} + \mathrm{Acc}_{\mathrm{novel}}}, \tag{17}$$

where $\mathrm{Acc}_{\mathrm{base}}$ and $\mathrm{Acc}_{\mathrm{novel}}$ are classification accuracies.

Cross-dataset evaluation trains on ImageNet and tests on other datasets without tuning, while domain generalization assesses robustness to distribution shift. Few-shot experiments follow standard 1/2/4/8/16-shot splits, with 16-shot subsets created by uniformly sampling 16 images per class and fixing the seed for fair comparison.

4.3. Implementation

We adhere to identical few-shot sampling, augmentations and backbones to ensure parity. EVR integrates the VSCR module described in Section 3.

4.4. Base-to-novel generalization (Table 1)

Table 1 shows that EVR consistently improves base-class accuracy while maintaining strong generalization to novel classes across diverse domains without textual supervision. Its VSCR module enhances harmonic mean by constructing compact prototype manifolds and balancing propagation with separation.

4.5. Cross-dataset transfer (Table 2)

Table 2 summarizes the results of cross-dataset transfer, where models are trained on ImageNet under a few-shot setting and directly evaluated on a range of other datasets. EVR demonstrates consistent improvements over existing methods, showing strong generalization across both object-centric and texture or scene-oriented datasets. These results highlight EVR's robustness in handling domain shifts and its effectiveness in transferring knowledge to unseen distributions.

4.6. Few-shot learning

Figure 5 shows that EVR achieves the highest average accuracy across 1–16 shots over 11 datasets. Performance gains grow with more shots, thanks to visually grounded prototypes and stable optimization.

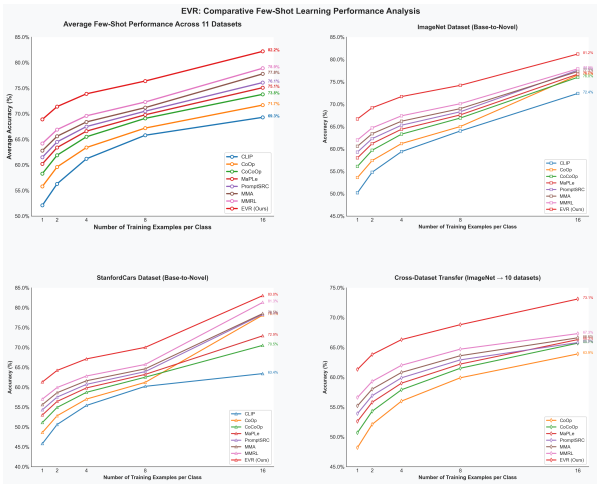

Figure 5. Average few-shot accuracy across 11 datasets (1/2/4/8/16 shots).

4.7. Comparison with Self-Supervised Learning Baselines

To verify that EVR gains are not simply inherited from CLIP's language supervision, we compare it with four recent *vision-only* self-supervised methods under the 1-shot / 16-shot protocol.

4.8. Ablation studies

We perform targeted ablations to quantify the contribution of key EVR components, reporting averages over 11 base-to-novel datasets unless otherwise stated. Table 4 compares variants with individual components removed and analyzes sensitivity to prototype count K and dimension d_p. Removing visual prototypes or graph-smoothing degrades both base and HM metrics, confirming the importance of EVR's design. Selected rows also report results on ImageNett and StanfordCars to illustrate per-dataset behavior.

Table 1. Base-to-novel generalization across 11 datasets. Each triple reports (Base / Novel / HM). Bold numbers indicate the best method in the column.

Method	Average			ImageNet[8]			Caltech101[13]			OxfordPets[41]		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP[44]	69.34	74.22	71.70	72.43	68.14	70.22	96.84	94.00	95.40	91.17	97.26	94.12
COMMA[21]	82.42	75.87	79.04	76.04	70.89	73.86	97.94	94.56	96.50	95.62	97.84	96.72
CoOp[80]	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
CoCoOp[79]	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43
ProDA[35]	81.56	72.30	76.65	75.40	70.23	72.72	98.27	93.23	95.68	95.43	97.83	96.62
KgCoOp[64]	80.73	73.60	77.00	75.83	69.96	72.78	97.72	94.39	96.03	94.65	97.76	96.18
MaPLe[23]	82.28	75.14	78.55	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58
MaPLe + LatHAdapter[75]	82.42	76.13	79.15	76.99	70.49	73.60	98.15	94.87	96.48	96.07	98.04	97.05
PromptSRC[24]	84.26	76.10	79.97	77.60	70.73	74.01	98.10	94.03	96.02	95.33	97.30	96.30
PromptKD[30]	84.15	78.98	81.48	77.20	70.90	73.92	98.50	96.90	97.69	94.50	96.80	95.64
ProVP[60]	85.20	73.22	78.76	75.82	69.21	72.36	98.92	94.21	96.51	95.87	97.65	96.75
MetaPrompt[76]	83.65	75.48	79.09	77.52	70.83	74.02	98.13	94.58	96.32	95.53	97.00	96.26
TCP[65]	84.13	75.36	79.51	77.27	69.87	73.38	98.23	94.67	96.42	94.67	97.20	95.92
MMA[63]	83.20	76.80	79.87	77.31	71.00	74.02	98.40	94.00	96.15	95.40	98.07	96.72
MMRL[15]	85.68	77.16	81.20	77.90	71.30	74.45	98.97	94.50	96.68	95.90	97.60	96.74
MMRL++[16]	85.53	78.32	81.77	77.63	71.50	74.44	99.07	94.53	96.75	95.60	97.43	96.51
ANPrompt[14]	86.15	77.70	81.70	77.83	71.17	74.35	98.97	94.73	96.80	95.73	97.17	96.44
EVR (Ours)	**89.12**	**80.59**	**84.63**	**81.20**	**74.60**	**77.70**	**99.30**	**95.10**	**97.15**	**96.50**	**98.20**	**97.34**

Method	StanfordCars[25] B / N / HM	Flowers102[39] B / N / HM	Food101[2] B / N / HM	FGVCAircraft[37] B / N / HM	SUN397[59] B / N / HM	DTD[6] B / N / HM	EuroSAT[18] B / N / HM	UCF101[39] B / N / HM
CLIP[44]	63.37 / 74.89 / 68.65	72.08 / 77.80 / 74.83	90.10 / 91.22 / 90.66	27.19 / 36.29 / 31.09	69.36 / 75.35 / 72.23	53.24 / 59.90 / 56.37	56.48 / 64.05 / 60.03	70.53 / 77.50 / 73.85
COMMA[21]	73.48 / 74.91 / 73.96	94.86 / 75.13 / 83.88	90.42 / 92.74 / 91.84	36.47 / 34.23 / 35.84	80.94 / 79.32 / 80.86	81.04 / 58.62 / 68.32	93.56 / 74.26 / 83.42	84.06 / 80.56 / 81.84
CoOp[80]	78.12 / 60.40 / 68.13	97.60 / 59.67 / 74.06	88.33 / 82.26 / 85.19	40.44 / 22.30 / 28.75	80.60 / 65.89 / 72.51	79.44 / 41.18 / 54.24	92.19 / 54.74 / 68.69	84.69 / 56.05 / 67.46
CoCoOp[79]	70.49 / 73.59 / 72.01	94.87 / 71.75 / 81.71	90.70 / 91.29 / 90.99	33.41 / 23.71 / 27.74	79.74 / 76.86 / 78.27	77.01 / 56.00 / 64.85	87.49 / 60.04 / 71.21	82.33 / 73.45 / 77.64
ProDA[35]	74.70 / 71.20 / 72.91	97.70 / 68.68 / 80.66	90.30 / 88.57 / 89.43	36.90 / 34.13 / 35.46	78.67 / 76.93 / 77.79	80.67 / 56.48 / 66.44	83.90 / 66.00 / 73.88	85.23 / 71.97 / 78.04
KgCoOp[64]	71.76 / 75.04 / 73.36	95.00 / 74.73 / 83.65	90.50 / 91.70 / 91.09	36.21 / 33.55 / 34.83	80.29 / 76.53 / 78.36	77.55 / 54.99 / 64.35	85.64 / 64.34 / 73.48	82.89 / 76.67 / 79.65
MaPLe[23]	72.94 / 74.00 / 73.47	95.92 / 72.46 / 82.56	90.71 / 92.05 / 91.38	37.44 / 35.61 / 36.50	80.82 / 78.70 / 79.75	80.36 / 59.18 / 68.16	94.07 / 73.23 / 82.35	83.00 / 78.66 / 80.77
MaPLe + LatHAdapter[75]	73.06 / 73.93 / 73.49	96.04 / 74.44 / 83.87	90.85 / 91.90 / 91.37	37.49 / 35.99 / 36.73	80.95 / 78.74 / 79.83	79.63 / 61.96 / 69.69	93.66 / 77.69 / 84.93	83.73 / 79.38 / 81.50
PromptSRC[24]	78.27 / 74.97 / 76.58	98.07 / 76.50 / 85.95	90.67 / 91.53 / 91.10	42.73 / 37.87 / 40.15	82.67 / 78.47 / 80.52	83.37 / 62.97 / 71.75	92.90 / 73.90 / 82.32	87.10 / 78.80 / 82.74
PromptKD[30]	80.60 / 82.40 / 81.49	98.80 / 82.00 / 89.62	89.50 / 91.70 / 90.59	45.70 / 44.10 / 44.88	83.20 / 80.30 / 81.72	82.40 / 69.10 / 75.16	87.20 / 74.20 / 80.17	88.10 / 80.40 / 84.07
PromptKD + LatHAdapter[75]	80.30 / 82.40 / 81.34	98.90 / 82.80 / 90.14	89.50 / 91.50 / 90.49	46.20 / 40.90 / 43.39	82.80 / 81.00 / 81.89	82.90 / 72.10 / 77.12	95.40 / 81.40 / 87.85	87.80 / 81.10 / 84.32
ProVP[60]	80.43 / 67.96 / 73.67	98.42 / 72.06 / 83.20	90.32 / 90.91 / 90.61	47.08 / 29.87 / 36.55	80.67 / 76.11 / 78.32	83.95 / 59.06 / 69.34	97.12 / 72.91 / 83.29	88.56 / 75.55 / 81.54
MetaPrompt[76]	76.34 / 75.01 / 75.48	97.66 / 74.49 / 84.52	90.74 / 91.85 / 91.29	40.14 / 36.51 / 38.24	82.26 / 79.04 / 80.62	83.10 / 58.05 / 68.35	93.53 / 75.21 / 83.38	85.33 / 77.72 / 81.35
TCP[65]	80.80 / 74.13 / 77.32	97.73 / 75.57 / 85.23	90.57 / 91.37 / 90.97	41.97 / 34.43 / 37.83	82.63 / 78.20 / 80.35	82.77 / 58.07 / 68.25	91.63 / 74.73 / 82.32	87.13 / 80.77 / 83.83
MMA[63]	78.50 / 73.10 / 75.70	97.77 / 75.93 / 85.48	90.13 / 91.30 / 90.71	40.57 / 36.33 / 38.33	82.27 / 78.57 / 80.38	83.20 / 65.63 / 73.38	85.46 / 82.34 / 83.87	86.23 / 80.03 / 82.20
MMRL[15]	81.30 / 75.07 / 78.06	98.97 / 77.27 / 86.78	90.57 / 91.50 / 91.03	46.30 / 37.03 / 41.15	83.20 / 79.30 / 81.20	85.67 / 65.00 / 73.82	95.60 / 80.17 / 87.21	88.10 / 80.07 / 83.89
MMRL++[16]	81.33 / 75.27 / 78.18	98.53 / 77.90 / 87.01	90.47 / 91.73 / 91.10	46.40 / 38.77 / 42.24	83.03 / 79.60 / 81.28	85.47 / 65.97 / 74.46	95.93 / 88.27 / 91.94	87.37 / 80.53 / 83.81
ANPrompt[14]	83.57 / 74.63 / 78.85	98.60 / 77.30 / 86.66	90.63 / 91.50 / 91.06	49.67 / 36.60 / 42.14	83.07 / 79.07 / 81.02	85.20 / 65.10 / 73.80	95.53 / 87.33 / 91.21	88.80 / 80.07 / 84.21
EVR (Ours)	**83.00 / 76.80 / 79.78**	**99.00 / 78.10 / 87.13**	**97.20 / 92.00 / 94.49**	**49.00 / 39.50 / 44.00**	**88.50 / 81.79 / 86.11**	**87.00 / 82.00 / 84.49**	**97.00 / 82.00 / 89.23**	**92.62 / 86.40 / 86.66**

Table 2. Cross-dataset evaluation. Models are trained on ImageNet (few-shot) and tested on target datasets without further tuning. Each column lists per-target accuracy; "Avg" is the average over the ten target datasets.

	Source	Target										
Method	ImageNet[8]	Avg	Caltech101[13]	OxfordPets[41]	StanfordCars[25]	Flowers102[39]	Food101[2]	FGVCAircraft[37]	SUN397[59]	DTD[6]	EuroSAT[18]	UCF101[39]
CoOp[80]	71.51	63.88	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55
CoCoOp[79]	71.02	65.74	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21
MaPLe[23]	70.72	66.30	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69
PromptSRC[24]	71.27	65.81	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75
TCP[65]	71.40	66.29	93.97	91.25	64.69	71.21	86.69	23.45	67.15	44.35	51.45	68.73
MMA[63]	71.00	66.61	93.80	90.30	66.13	72.07	86.12	25.33	68.17	46.57	49.24	68.32
SurPL-G[31]	73.33	66.61	93.73	90.16	64.67	70.53	85.52	24.80	67.43	48.54	52.85	67.86
BIP-D[66]	70.83	66.57	93.93	90.13	65.53	71.43	86.27	26.47	67.23	46.00	48.90	69.83
MMRL[15]	72.03	67.25	94.67	91.43	66.10	72.77	86.40	26.30	67.57	45.90	53.10	68.27
MMRL++[16]	71.87	69.03	94.63	91.43	66.60	73.53	86.73	26.07	67.77	46.13	53.00	69.03
ANPrompt[14]	71.13	67.14	94.70	91.00	66.00	72.87	86.23	26.10	67.47	45.83	51.97	69.27
PromptSRC + LatHAdapter[75]	71.77	66.40	93.87	89.92	64.86	70.65	86.22	24.30	66.79	48.46	51.04	67.88
EVR (Ours)	**73.06**	**69.38**	**95.20**	**92.80**	**67.50**	**73.80**	**89.23**	**29.60**	**68.81**	**52.52**	**54.24**	**70.11**

4.9. Domain generalization (Table 5)

Table 5 reports the results of domain generalization, where models are trained on ImageNet and evaluated on several domain-shifted variants. EVR demonstrates superior robustness compared to previous methods, consistently achieving strong performance across multiple challenging benchmarks. These results highlight EVR's ability to generalize effectively under domain shift conditions.

Table 3. Comparison of self-supervised learning methods and proposed approaches on datasets (1-shot / 16-shot settings).

Method	ImageNet[8]	Aircraft[37]	Flowers[39]	EuroSAT[18]	Avg Acc
DINOv2[40]	71.1 / 73.9	28.9 / 50.6	85.2 / 97.5	67.8 / 86.8	63.3 / 77.2
SynCLR[52]	71.5 / 74.0	27.4 / 51.0	84.4 / 97.7	74.4 / 86.9	64.4 / 77.4
MAE[17]	71.2 / 74.3	26.4 / 47.6	83.6 / 97.4	66.5 / 86.7	61.9 / 76.3
HoM-DINO[62]	71.5 / 74.3	29.4 / 51.2	86.3 / 98.3	74.5 / 87.8	65.4 / 77.9
EVR (Ours)	**74.2 / 77.0**	**49.0 / 79.5**	**95.0 / 99.8**	**87.0 / 97.0**	**76.3 / 88.3**

Table 4. Ablation results. All numbers are mean ± std over 3 seeds (average across 11 datasets) unless noted.

Variant	Base (%)	Novel (%)	HM (%)
Full EVR	**89.12 ± 0.23**	**80.59 ± 0.31**	**84.63 ± 0.27**
w/o V (no visual prototypes)	85.50 ± 0.41	78.20 ± 0.38	81.70 ± 0.39
w/o GS (no graph smoothing)	86.00 ± 0.35	78.50 ± 0.42	82.10 ± 0.37
w/o Separation	87.00 ± 0.29	79.00 ± 0.45	82.80 ± 0.36
w/o Entropy ($\mathcal{L}_{\text{ent}}$)	87.85 ± 0.32	79.35 ± 0.39	83.40 ± 0.34
w/o EMA (no prototype EMA)	86.75 ± 0.47	78.90 ± 0.51	82.60 ± 0.48
w/o Adaptive Loss Norm	87.40 ± 0.38	79.20 ± 0.43	83.10 ± 0.40
detach assignments (prototypes detached in Eq. 4)	88.25 ± 0.33	79.85 ± 0.36	83.90 ± 0.34
w/o Propagation (W_g removed)	85.90 ± 0.44	78.35 ± 0.49	81.95 ± 0.46
clustered-buffer (alt. reinit strategy)	88.65 ± 0.28	80.10 ± 0.34	84.20 ± 0.30
Prototype count K (sensitivity)			
$K = 32$	86.50 ± 0.39	79.00 ± 0.44	82.60 ± 0.41
$K = 128$	87.20 ± 0.34	79.50 ± 0.41	83.20 ± 0.37
$K = 512$	89.12 ± 0.23	80.59 ± 0.31	84.63 ± 0.27
$K = 2048$	88.35 ± 0.31	80.25 ± 0.38	84.20 ± 0.34

Variant	ImageNet[8]			StanfordCars[25]		
	Base	Novel	HM	Base	Novel	HM
Full EVR	81.20 ± 0.15	74.60 ± 0.18	77.70 ± 0.16	83.00 ± 0.12	76.80 ± 0.15	79.78 ± 0.13
w/o EMA	80.50 ± 0.20	73.80 ± 0.22	77.00 ± 0.21	82.40 ± 0.14	76.20 ± 0.17	79.10 ± 0.15
w/o Adaptive Norm	80.80 ± 0.17	74.20 ± 0.19	77.30 ± 0.18	82.60 ± 0.15	76.40 ± 0.17	79.35 ± 0.16
detach assignments	81.00 ± 0.16	74.40 ± 0.19	77.50 ± 0.17	82.80 ± 0.14	76.60 ± 0.16	79.55 ± 0.15
w/o Entropy	80.70 ± 0.18	74.20 ± 0.20	77.30 ± 0.19	82.40 ± 0.16	76.20 ± 0.18	79.20 ± 0.17

Prototype dimension d_p (avg over 11 datasets)

d_p	Base (%)	Novel (%)	HM (%)
32	86.50	79.00	82.60
128	87.20	79.50	83.20
256	88.00	80.00	83.80
512	**89.12**	**80.59**	**84.63**
1024	88.80 ± 0.28	80.40 ± 0.35	84.45 ± 0.31

Table 5. Domain generalization. Models are trained on ImageNet and evaluated on four domain-shifted variants.

Method	ImageNet-V2[46]	ImageNet-Sketch[54]	ImageNet-A[20]	ImageNet-R[19]	Avg
CLIP[44]	60.83	46.15	47.77	73.96	57.18
CoOp[80]	64.20	47.99	49.71	75.21	59.28
CoCoOp[79]	64.07	48.75	50.63	76.18	59.90
MaPLe[23]	64.07	49.15	50.90	76.98	60.26
PromptSRC[24]	64.35	49.55	50.90	77.80	60.63
HiCroPL[77]	64.33	49.47	50.79	77.15	60.44
ANPrompt[14]	64.63	49.13	50.37	77.47	60.40
MMA[63]	64.33	49.13	51.12	77.32	60.48
MMRL[15]	64.47	49.17	51.20	77.53	60.59
MMRL++[16]	64.67	49.30	51.00	77.43	60.60
EVR (Ours)	**65.00**	**49.80**	**51.70**	**78.00**	**61.13**

4.10. Hyperparameter sweeps

We sweep the balance weight α and the penalty coefficient λ. Table 6 shows $\alpha = 0.7$ and $\lambda = 0.5$ strike the best trade-off between adaptation and generalization.

Table 6. Comprehensive hyperparameter sensitivity analysis (averaged over 11 datasets), with bold indicating default settings that yield optimal harmonic mean (HM) of base and novel accuracies.

Hyperparameter	Value	Base (%)	Novel (%)	HM (%)
α	0.0	85.00	78.00	81.30
	0.3	87.50	79.20	83.10
	0.5	88.20	79.80	83.80
	0.7	**89.12**	**80.59**	**84.63**
	1.0	87.80	79.50	83.40
λ	0.0	86.00	79.00	82.30
	0.2	87.00	79.50	83.00
	0.5	**89.12**	**80.59**	**84.63**
	2.0	88.50	80.00	84.10
	4.0	87.20	79.30	83.10
τ	0.05	88.80	80.20	84.30
	0.1	89.00	80.40	84.50
	0.2	**89.12**	**80.59**	**84.63**
	0.5	88.60	80.10	84.20
α_{ema}	0.9	88.70	80.30	84.40
	0.99	89.00	80.50	84.60
	0.996	**89.12**	**80.59**	**84.63**
	0.999	88.90	80.40	84.50
r_p	0.3	88.80	80.20	84.30
	0.5	**89.12**	**80.59**	**84.63**
	0.7	89.00	80.50	84.60
	1.0	88.60	80.10	84.20

5. Conclusion

EVR is a vision-only framework that learns structured semantic manifolds directly from images, without textual supervision. It discovers multiple visual prototypes via self-supervised clustering, organizes them into a graph to model semantic relations, and jointly optimizes embeddings and prototypes through contrastive consistency and propagation. This design emphasizes visual signals, mitigates early language bias, and captures intra-class diversity. Experiments show EVR consistently improves transfer learning across base-to-novel generalization, cross-dataset evaluation, domain robustness, and few-shot tasks. These gains stem from the VSCR module's visually grounded prototypes and stable training. Ablations validate the roles of multi-prototype modeling and adaptive loss normalization, while visualizations highlight improved class separation. Despite its effectiveness, EVR requires strong visual pretraining and introduces graph-related overhead. Overall, EVR offers a robust alternative to language-centric pipelines by learning a purely visual semantic manifold. Future work will explore efficient propagation, scalable graph maintenance, and optional textual integration for interpretability.

References

[1] Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to dino: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22025–22035, 2025. 1

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 5, 7

[3] 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. 1

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[5] Eulrang Cho, Jooyeon Kim, and Hyunwoo J Kim. Distribution-aware prompt tuning for vision-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22004–22013, 2023. 2

[6] 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*, pages 3606–3613, 2014. 5, 7

[7] 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. 2

[8] 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*, pages 248–255. Ieee, 2009. 5, 7, 8

[9] Xing Di, Yiyu Zheng, Xiaoming Liu, and Yu Cheng. Pros: Facial omni-representation learning via prototype-based self-distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6087–6098, 2024. 1

[10] Aniket Didolkar, Andrii Zadaianchuk, Rabiul Awal, Maximilian Seitzer, Efstratios Gavves, and Aishwarya Agrawal. Ctrl-o: language-controllable object-centric visual representation learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29523–29533, 2025. 1

[11] Hao Dong, Lijun Sheng, Jian Liang, Ran He, Eleni Chatzi, and Olga Fink. Adapting vision-language models without labels: A comprehensive survey. *arXiv preprint arXiv:2508.05547*, 2025. 2

[12] Matteo Farina, Massimiliano Mancini, Giovanni Iacca, and Elisa Ricci. Rethinking few-shot adaptation of vision-language models in two stages. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29989–29998, 2025. 2

[13] 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*, pages 178–178. IEEE, 2004. 5, 7

[14] Yansheng Gao, Yufei Zheng, Jinghan Qu, Zixi Zhu, Yukuan Zhang, and Shengsheng Wang. Anprompt: Anti-noise prompt tuning for vision-language models. *arXiv preprint arXiv:2508.04677*, 2025. 7, 8

[15] Yuncheng Guo and Xiaodong Gu. Mmrl: Multi-modal representation learning for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25015–25025, 2025. 7, 8

[16] Yuncheng Guo and Xiaodong Gu. Mmrl++: Parameter-efficient and interaction-aware representation learning for vision-language models. *arXiv preprint arXiv:2505.10088*, 2025. 7, 8

[17] 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*, pages 16000–16009, 2022. 8

[18] 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. 5, 7, 8

[19] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 5, 8

[20] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021. 5, 8

[21] Lianyu Hu, Liqing Gao, Zekang Liu, Chi-Man Pun, and Wei Feng. Comma: Co-articulated multi-modal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2238–2246, 2024. 7

[22] Licheng Jiao, Jie Chen, Fang Liu, Shuyuan Yang, Chao You, Xu Liu, Lingling Li, and Biao Hou. Graph representation learning meets computer vision: A survey. *IEEE Transactions on Artificial Intelligence*, 4(1):2–22, 2022. 2

[23] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023. 7, 8

[24] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of*

the IEEE/CVF international conference on computer vision, pages 15190–15200, 2023. 7, 8

[25] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 5, 7, 8

[26] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020. 1

[27] Ruiyu Li, Makarand Tapaswi, Renjie Liao, Jiaya Jia, Raquel Urtasun, and Sanja Fidler. Situation recognition with graph neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4173–4182, 2017. 2

[28] Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. *Advances in Neural Information Processing Systems*, 36:13448–13466, 2023. 2

[29] Zhiyuan Li, Srinadh Bhojanapalli, Manzil Zaheer, Sashank Reddi, and Sanjiv Kumar. Robust training of neural networks using scale invariant architectures. In *International Conference on Machine Learning*, pages 12656–12684. PMLR, 2022. 2

[30] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26617–26626, 2024. 7

[31] Liangchen Liu, Nannan Wang, Xi Yang, Xinbo Gao, and Tongliang Liu. Surrogate prompt learning: Towards efficient and diverse prompt learning for vision-language models. In *Forty-second International Conference on Machine Learning*. 7

[32] Wenzhuo Liu, Xin-Jian Wu, Fei Zhu, Ming-Ming Yu, Chuang Wang, and Cheng-Lin Liu. Class incremental learning with self-supervised pre-training and prototype learning. *Pattern Recognition*, 157:110943, 2025. 2

[33] Yixin Liu, Yizhen Zheng, Daokun Zhang, Vincent CS Lee, and Shirui Pan. Beyond smoothing: Unsupervised graph representation learning with edge heterophily discriminating. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4516–4524, 2023. 2

[34] Bin Lu, Xiaoying Gan, Lina Yang, Weinan Zhang, Luoyi Fu, and Xinbing Wang. Geometer: Graph few-shot class-incremental learning via prototype representation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 1152–1161, 2022. 2

[35] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5206–5215, 2022. 7

[36] Chaofan Ma, Yuhuan Yang, Yanfeng Wang, Ya Zhang, and Weidi Xie. Open-vocabulary semantic segmentation with frozen vision-language models. *arXiv preprint arXiv:2210.15138*, 2022. 1

[37] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5, 7, 8

[38] Ans Munir, Faisal Z Qureshi, Muhammad Haris Khan, and Mohsen Ali. Tlac: Two-stage lmm augmented clip for zero-shot classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 159–169, 2025. 2

[39] 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*, pages 722–729. IEEE, 2008. 5, 7, 8

[40] 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. 8

[41] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 5, 7

[42] Dingkang Peng, Xiaokang Zhang, Wanjing Wu, Xianping Ma, and Weikang Yu. Osclip: Domain-adaptive prompt tuning of vision-language models for open-set remote sensing image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025. 2

[43] Fang Peng, Xiaoshan Yang, Linhui Xiao, Yaowei Wang, and Changsheng Xu. Sgva-clip: Semantic-guided visual adapting of vision-language models for few-shot image classification. *IEEE Transactions on Multimedia*, 26:3469–3480, 2023. 2

[44] 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*, pages 8748–8763. PmLR, 2021. 7, 8

[45] Alessio Ragno, Biagio La Rosa, and Roberto Capobianco. Prototype-based interpretable graph neural networks. *IEEE Transactions on Artificial Intelligence*, 5(4):1486–1495, 2022. 2

[46] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 5, 8

[47] Fereshteh Shakeri, Yunshi Huang, Julio Silva-Rodríguez, Houda Bahig, An Tang, Jose Dolz, and Ismail Ben Ayed. Few-shot adaptation of medical vision-language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 553–563. Springer, 2024. 2

[48] Julio Silva-Rodriguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23681–23690, 2024. 2

[49] Lin Song, Ruoyi Xue, Hang Wang, Hongbin Sun, Yixiao Ge, Ying Shan, et al. Meta-adapter: An online few-shot learner for vision-language model. *Advances in Neural Information Processing Systems*, 36:55361–55374, 2023. 2

[50] Thomas Stegmüller, Tim Lebailly, Behzad Bozorgtabar, Tinne Tuytelaars, and Jean-Philippe Thiran. Croc: Cross-view online clustering for dense visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7000–7009, 2023. 1

[51] Baofeng Tan, Xiu-Shen Wei, and Lin Zhao. Prototype-based contrastive learning with stage-wise progressive augmentation for self-supervised fine-grained learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4125–4134, 2025. 1

[52] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15887–15898, 2024. 8

[53] Muhammad Atta ur Rahman, Dooseop Choi, Seung-Ik Lee, and KyoungWook Min. Beyond-labels: Advancing open-vocabulary segmentation with vision-language models. In *2025 17th International Conference on Advanced Computational Intelligence (ICACI)*, pages 231–236. IEEE, 2025. 1

[54] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019. 5, 8

[55] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6857–6866, 2018. 2

[56] Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. *Advances in neural information processing systems*, 35:16423–16438, 2022. 1

[57] Jiamin Wu, Tianzhu Zhang, Zheng-Jun Zha, Jiebo Luo, Yongdong Zhang, and Feng Wu. Prototype-augmented self-supervised generative network for generalized zero-shot learning. *IEEE Transactions on Image Processing*, 33:1938–1951, 2024. 2

[58] Likang Wu, Zhi Li, Hongke Zhao, Zhefeng Wang, Qi Liu, Baoxing Huai, Nicholas Jing Yuan, and Enhong Chen. Recognizing unseen objects via multimodal intensive knowledge graph propagation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2618–2628, 2023. 2

[59] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 5, 7

[60] Chen Xu, Yuhan Zhu, Haocheng Shen, Boheng Chen, Yixuan Liao, Xiaoxin Chen, and Limin Wang. Progressive visual prompt learning with contrastive feature re-formation. *International Journal of Computer Vision*, 133(2):511–526, 2025. 7

[61] Zhongxing Xu, Feilong Tang, Zhe Chen, Yingxue Su, Zhiyi Zhao, Ge Zhang, Jionglong Su, and Zongyuan Ge. Toward modality gap: Vision prototype learning for weakly-supervised semantic segmentation with clip. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9023–9031, 2025. 1

[62] Haoyuan Yang, Xiaoou Li, Jiaming Lv, Xianjun Cheng, Qilong Wang, and Peihua Li. Imaginefsl: Self-supervised pre-training matters on imagined base set for vlm-based few-shot learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30020–30031, 2025. 8

[63] Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23826–23837, 2024. 7, 8

[64] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6757–6767, 2023. 7

[65] Hantao Yao, Rui Zhang, and Changsheng Xu. Tcp: Textual-based class-aware prompt tuning for visual-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23438–23448, 2024. 7

[66] Hantao Yao, Rui Zhang, Huaihai Lyu, Yongdong Zhang, and Changsheng Xu. Bi-modality individual-aware prompt tuning for visual-language model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 7

[67] Siyuan Yao, Yang Guo, Yanyang Yan, Wenqi Ren, and Xiaochun Cao. Unctrack: Reliable visual object tracking with an uncertainty-aware prototype memory network. *IEEE Transactions on Image Processing*, 2025. 2

[68] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13834–13844, 2021. 2

[69] Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1593–1603, 2024. 2

[70] Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Online deep clustering for unsupervised representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6688–6697, 2020. 1

[71] Lu Zhang, Lu Qi, Xu Yang, Hong Qiao, Ming-Hsuan Yang, and Zhiyong Liu. Automatically discovering novel visual categories with self-supervised prototype learning. *arXiv preprint arXiv:2208.00979*, 2022. 1, 2

[72] Lu Zhang, Lu Qi, Xu Yang, Hong Qiao, Ming-Hsuan Yang, and Zhiyong Liu. Automatically discovering novel visual categories with adaptive prototype learning. *IEEE transactions on pattern analysis and machine intelligence*, 46(4): 2533–2544, 2023. 1

[73] Wei Zhang, Jingyang Qiao, Yuan Xie, Zhizhong Zhang, and Xin Tan. Efficient prototypical classifier for class-

incremental learning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2

[74] Yang Zhao, Hao Zhang, and Xiuyuan Hu. Penalizing gradient norm for efficiently improving generalization in deep learning. In *International conference on machine learning*, pages 26982–26992. PMLR, 2022. 2

[75] Yumiao Zhao, Bo Jiang, Yuhe Ding, Xiao Wang, Jin Tang, and Bin Luo. Fine-grained vlm fine-tuning via latent hierarchical adapter learning. *arXiv preprint arXiv:2508.11176*, 2025. 7

[76] Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, et al. Clip in medical imaging: A comprehensive survey. *arXiv preprint arXiv:2312.07353*, 2023. 7

[77] Hao Zheng, Shunzhi Yang, Zhuoxin He, Jinfeng Yang, and Zhenhua Huang. Hierarchical cross-modal prompt learning for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1891–1901, 2025. 8

[78] Yimei Zheng and Caiyan Jia. Protomgae: prototype-aware masked graph auto-encoder for graph representation learning. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–22, 2024. 1, 2

[79] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 7, 8

[80] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 7, 8

[81] Xingyu Zhu, Shuo Wang, Beier Zhu, Miaoge Li, Yunfan Li, Junfeng Fang, Zhicai Wang, Dongsheng Wang, and Hanwang Zhang. Dynamic multimodal prototype learning in vision-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2501–2511, 2025. 2